NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

THE USE OF TWITTER TO PREDICT THE LEVEL OF INFLUENZA ACTIVITY IN THE UNITED STATES

by

Kok Wah Ng

September 2014

Thesis Advisor:	Samuel E. Buttrey
Second Reader:	Nedialko Dimitrov

Approved for public release; distribution is unlimited

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704–0188

1. AGENCY USE ONLY *(Leave blank)*	**2. REPORT DATE** September 2014	**3. REPORT TYPE AND DATES COVERED** Master's Thesis

4. TITLE AND SUBTITLE
THE USE OF TWITTER TO PREDICT THE LEVEL OF INFLUENZA ACTIVITY IN THE UNITED STATES

5. FUNDING NUMBERS

6. AUTHOR(S) Kok Wah Ng

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Naval Postgraduate School
Monterey, CA 93943-5000

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING /MONITORING AGENCY NAME(S) AND ADDRESS(ES)
N/A

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government. IRB Protocol number ____N/A____.

12a. DISTRIBUTION / AVAILABILITY STATEMENT	**12b. DISTRIBUTION CODE**
Approved for public release; distribution is unlimited	A

13. ABSTRACT (maximum 200 words)

Controlling the outbreak of epidemic diseases such as influenza has always been a concern for the United States. Traditional surveillance tools such as the ILINet and Virologic provide the Centers for Disease Control and Prevention (CDC) with influenza surveillance statistics at a lag of 1 to 2 weeks. The CDC requires a tool that can forecast the level of influenza activity.

The rise in the popularity of social media websites such as Flickr, Twitter and Facebook has transformed the web into an interactive sharing platform. The huge amount of generated unstructured data has become an invaluable source for detecting patterns or novelties.

This research explores the correlation between Twitter messages (tweets) and CDC ILI and Virologic surveillance data. Using 17 months of tweets, regression models are developed to predict influenza-related statistics. The proposed approach aggregates the weekly frequencies of hand-chosen words that are indicative of an influenza attack using separate predictor variables. The predictions generated by the best models are found to have a Pearson's correlation coefficient of 0.900 (95% CI: 0.732, 0.965) and 0.833 (95% CI: 0.574, 0.940) against the CDC ILI surveillance data and CDC Virologic surveillance data, respectively.

14. SUBJECT TERMS correlation, data analysis, Twitter, tweet, influenza	**15. NUMBER OF PAGES** 125
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	**18. SECURITY CLASSIFICATION OF THIS PAGE**	**19. SECURITY CLASSIFICATION OF ABSTRACT**	**20. LIMITATION OF ABSTRACT**
Unclassified	Unclassified	Unclassified	UU

NSN 7540–01-280-5500 Standard Form 298 (Rev. 2–89)
Prescribed by ANSI Std. 239–18

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release; distribution is unlimited

THE USE OF TWITTER TO PREDICT THE LEVEL OF INFLUENZA ACTIVITY IN THE UNITED STATES

Kok Wah Ng
Civilian, Singapore Technologies Engineering
B.Eng., Nanyang Technological University, 2007

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

NAVAL POSTGRADUATE SCHOOL
September 2014

Author: Kok Wah Ng

Approved by: Samuel E. Buttrey
Thesis Advisor

Nedialko Dimitrov
Second Reader

Robert Dell, Ph.D.
Chair, Department of Operations Research

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

Controlling the outbreak of epidemic diseases such as influenza has always been a concern for the United States. Traditional surveillance tools such as the ILINet and Virologic provide the Centers for Disease Control and Prevention (CDC) with influenza surveillance statistics at a lag of 1 to 2 weeks. The CDC requires a tool that can forecast the level of influenza activity.

The rise in the popularity of social media websites such as Flickr, Twitter and Facebook has transformed the web into an interactive sharing platform. The huge amount of generated unstructured data has become an invaluable source for detecting patterns or novelties.

This research explores the correlation between Twitter messages (tweets) and CDC ILI and Virologic surveillance data. Using 17 months of tweets, regression models are developed to predict influenza-related statistics. The proposed approach aggregates the weekly frequencies of hand-chosen words that are indicative of an influenza attack using separate predictor variables. The predictions generated by the best models are found to have a Pearson's correlation coefficient of 0.900 (95% CI: 0.732, 0.965) and 0.833 (95% CI: 0.574, 0.940) against the CDC ILI surveillance data and CDC Virologic surveillance data, respectively.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF ACRONYMS AND ABBREVIATIONS

CDC	Centers for Disease Control and Prevention
GFT	Google Flu Trends
HHS	Health & Human Services
ILI	influenza-like illnesses
ILINet	U.S. Outpatient Influenza-like Illnesses Surveillance Network
JSON	JavaScript Object Notation
LASSO	Least Absolute Shrinkage and Selection Operator
NL	national level
NREVSS	National Respiratory and Enteric Virus Surveillance System
POS	part of speech
WHO	World Health Organization

THIS PAGE INTENTIONALLY LEFT BLANK

EXECUTIVE SUMMARY

In 2009, the emergence of the pandemic influenza type A (H1N1, also known as “swine flu”) in the U.S. and Mexico, led to a global spread that impacted world societies, economies and tourism. In order to prevent such influenza pandemics, U.S. health agencies need to be alerted of the danger ahead of time. The U.S. Centers for Disease Control and Prevention (CDC) currently relies on traditional influenza surveillance tools such as the Outpatient Influenza-Like Illnesses Surveillance Network (ILINet) and the Influenza Virologic Surveillance Network that track the number of outpatient ILI visits and the number of respiratory specimens collected and tested positive. These networks typically report at a lag of 1 to 2 weeks. Hence, it is evident that the CDC require tools that are able to predict the level of influenza activities, so as to be able to respond promptly and accurately.

In recent years, there has been a rise in the popularity of social media websites such as Flickr, YouTube, Twitter, and Facebook. Social media has transformed the web into an interactive sharing platform where huge amounts of unstructured data are generated every minute. The data from these sites has become an invaluable resource for researchers to detect patterns or novelties. Researchers have used real-time tweets to perform sentiment analysis to gauge public opinions as well as to explore the use of tweets to gain situational awareness of major events such as a snowstorm disaster. Both Culotta and Kim et al. have attempted to develop regression models using individual influenza-related keyword frequencies as predictor variables, to predict the level of influenza activity in the U.S. and Korea, respectively.

Similarly, this research attempts to develop and evaluate regression model(s) to predict influenza-related statistics such as the number of outpatient ILI visits and the number of respiratory specimens collected and tested positive. In contrast to Culotta and Kim et al., this research explores the method of aggregating frequencies of categories of hand-chosen terms as predictor variables. The proposed keyword categories include flu symptoms, flu activities, rest activities, flu-related verbs and adjectives, as well as emoticons expressing sadness. The mention of the term “flu” in a tweet may or may not

indicate a flu attack. The co-existence of other flu-related terms, however, can further strengthen and support the claim that an influenza-related event exists.

Regression analysis is then performed using the weekly time series dataset (populated with the predictor and response variables) to find a best subset of predictor variables to construct the model. Variation selection techniques such as exhaustive search and cross-validation are used to identify the best subset of predictor variables.

The resulting models for the national level seem to suggest the presence of correlation between the tweets and the CDC traditional surveillance data. The models give a fairly good prediction, capturing the increasing and declining trend throughout the U.S. flu season. For the model predicting the number of outpatient ILI visits, the Pearson's correlation coefficient between the test set predictions of model and actual CDC ILI surveillance data is computed to be 0.900 (95% CI: 0.732, 0.965). Similarly, for the model predicting the number of respiratory specimens collected, the Pearson's correlation coefficient between the test set predictions and actual CDC virologic surveillance data is computed to be 0.833 (95% CI: 0.574, 0.940).

The research has strengthened the claim that Twitter is a potential solution to the CDC's need for an early indicator of influenza activity level. The exploration of using aggregated frequencies of keyword categories as predictor variables seems to be successful. At the national level, the constructed models are able to provide a good weekly estimate of influenza activity indicators such as number of ILI outpatient visits and number of collected respiratory specimens.

ACKNOWLEDGMENTS

I would like to thank my family members for their constant support and assistance throughout my stay in Naval Postgraduate School. My wife, Jamie, has constantly offered motivating advice as well as reminders of my strengths.

It is with immense gratitude that I acknowledge the support and help of Associate Professor Samuel E. Buttrey and Assistant Professor Nedialko Dimitrov. They have always been patient, sincere, encouraging, and helpful throughout the period of this research.

I share the credit of my work with Captain Ittai Bar-Ilan from the Israeli Defense Force. I have thoroughly enjoyed his company as a friend, a classmate, and a teammate.

Lastly, a big thanks to my boss, Mr. Wong Fook Hoi, for encouraging me to sign up for this rewarding postgraduate program.

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

The United States has faced the challenge of controlling the highly contagious flu for decades. Each year, campaigns have been run to educate, remind, and encourage citizens to get the flu vaccine. Such campaigns have certainly helped to reduce the number of influenza-like illnesses (ILI).

In 2009, the emergence of the pandemic influenza type A (H1N1, also known as "swine flu") in the U.S. and Mexico led to a global spread (CDC 2010), at which time the United States was hit with 12,469 deaths (CDC 2014). The declaration of failure to contain the spread of influenza further led to precautionary steps taken by countries to control the infection. Employers were hit economically due to temporary closures of workplaces. Hundreds of schools were temporarily closed after seeing a jump in the number of students diagnosed with flu (Babwin 2009). Business and holiday travel plans were also impacted by the spread. Various countries took up precautionary measures such as quarantining visitors or citizens who were returning from a flu-infected area.

It is evident that the U.S. public health agencies must have good situational awareness to respond promptly and accurately to prevent the spread of such influenza pandemics. This situational awareness is often provided by early indicators of influenza-related activities, such as the number of influenza-associated hospitalizations, the number of outpatient visits associated with influenza, and the number of samples collected and positively tested. The development and usage of tools for predicting the level of influenza-related activities in a given geographic region can allow hospitals and clinics to prepare for inflows of patients, as well as logistics time for the distributions of antivirals and vaccines.

In recent years, there has been a rise in the popularity of social media websites such as Flickr, YouTube, Twitter, and Facebook, which have impacted the lives of people. These have provided an avenue for people, organizations, and even countries to conveniently interact and share information across the world. Social media has transformed the web into an interactive sharing platform where huge amounts of

unstructured data are generated every minute. The data from these sites has become an invaluable resource for detecting patterns or novelties.

Twitter has 271 million monthly active users, and 500 million Twitter messages (tweets) sent each day (Twitter 2014). Each tweet can contain text messages, shared images, and links to videos. Twitter users share their opinions on various subjects, as well as information pertaining to their well-being, locations, and plans.

This study evaluates the use of tweets to predict the level of influenza-related activity in the U.S. at the national level, the regional level, and the state level. Influenza-related tweets are first filtered from large databases of tweets. Regression analyses are then performed to determine if there is a relationship between the influenza-related tweets and the actual influenza activity data collected by the Centers for Disease Control and Prevention (CDC). This approach aggregates the frequencies of categories of hand-chosen terms of the tweets and CDC's weekly influenza-related statistics into a time series dataset. Prediction models are then constructed using a training set (a subset of the dataset). The rest of the dataset are then used as an independent test set to validate the prediction models. The presence of correlation between the test set predictions and actual CDC surveillance data would support the idea of using tweets as a leading indicator of influenza activities in the U.S.

II. BACKGROUND

A. INFLUENZA

Influenza is a viral infection that affects the well-being of people. Symptoms such as high fever, aching muscles, and headache can stay with the carrier for about a week. The spread of influenza from a carrier to an uninfected person occurs via particles released from a carrier's cough or sneeze. Flu is caused by an infection of influenza virus type A, B, or C. Seasonal outbreaks of flu are typically caused by influenza types A or B. In 2009, the emergence of a new influenza, type A H1N1, in the United States and Mexico resulted in a global spread of influenza. CDC estimates that a total of 60 million cases and a death count of 12,469 are attributable to H1N1 (CDC 2014).

1. Preventive Measures

Public campaigns are run annually to promote awareness and to educate people on the severity of flu infection. These campaigns encourage people to get a flu vaccine to protect them from the flu. The CDC has reported that flu vaccinations can reduce the risk of serious flu outcomes, such as hospitalization and death. A recent study shows that flu vaccination is associated with a 71% reduction in influenza-related hospitalizations (CDC 2014).

2. Influenza in U.S.

In the U.S., the flu season starts as early as October and ends as late as May of the following year. Figure 1 shows the number of outpatient visits associated with influenza-like illnesses (ILI) for the 2012–13 season. According to the CDC's classification, outpatient visits are considered ILI if the patient is diagnosed with a high fever of 100°F and cough or sore throat. The frequency of visits begins to increase in early October, peaks in January, and then declines.

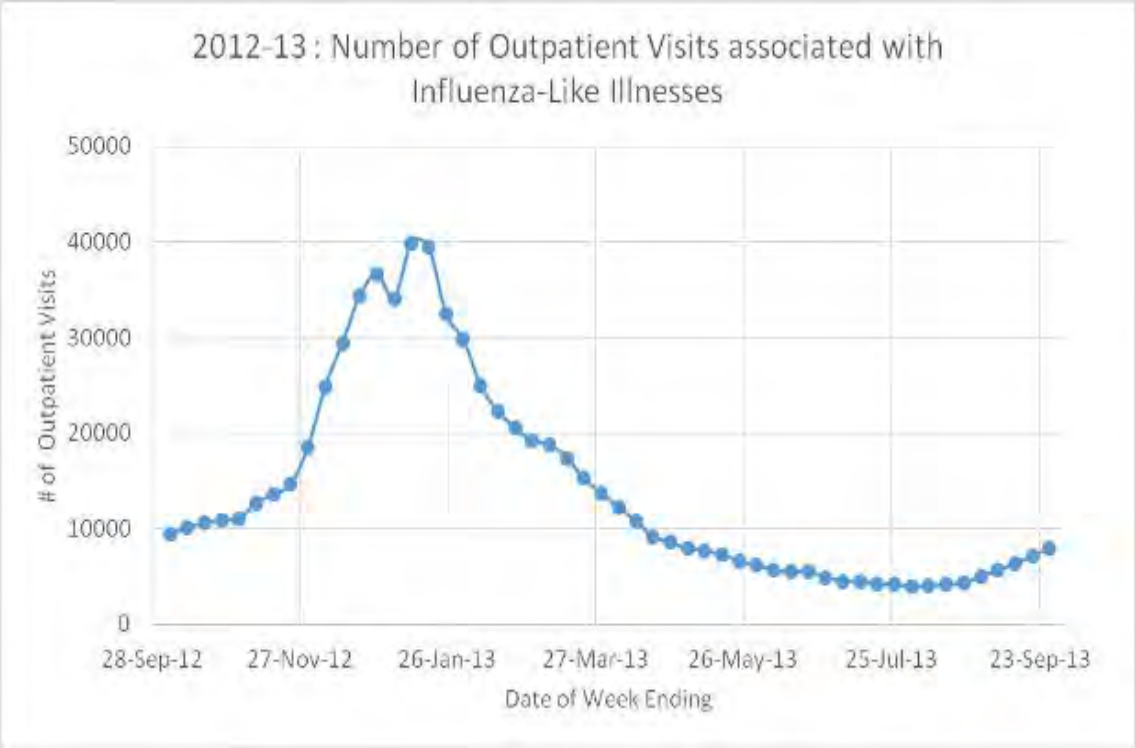

Figure 1. Number of Outpatient Visits Associated with Influenza-Like Illnesses (from CDC 2014)

January is regarded as the month in which the U.S. usually sees the peak in the level of influenza activity, and is associated with the most severe level of geographic spread. The geographic spread is a measure of the number of areas within a state that are seeing influenza activity. In January 2014, the geographic spread of influenza in most of the states reached the "widespread" level according to the CDC's categorization. Figure 2 shows that for the week ending January 18, 2014, 90% of the country is reported to be at the widespread level of the geographic spread.

In the 2012–2013 flu season, the CDC reported 728,957 ILI-related outpatient visits at the national level. Of the samples collected from these visits, 75,333 tested positive for influenza type A or B (CDC 2014).

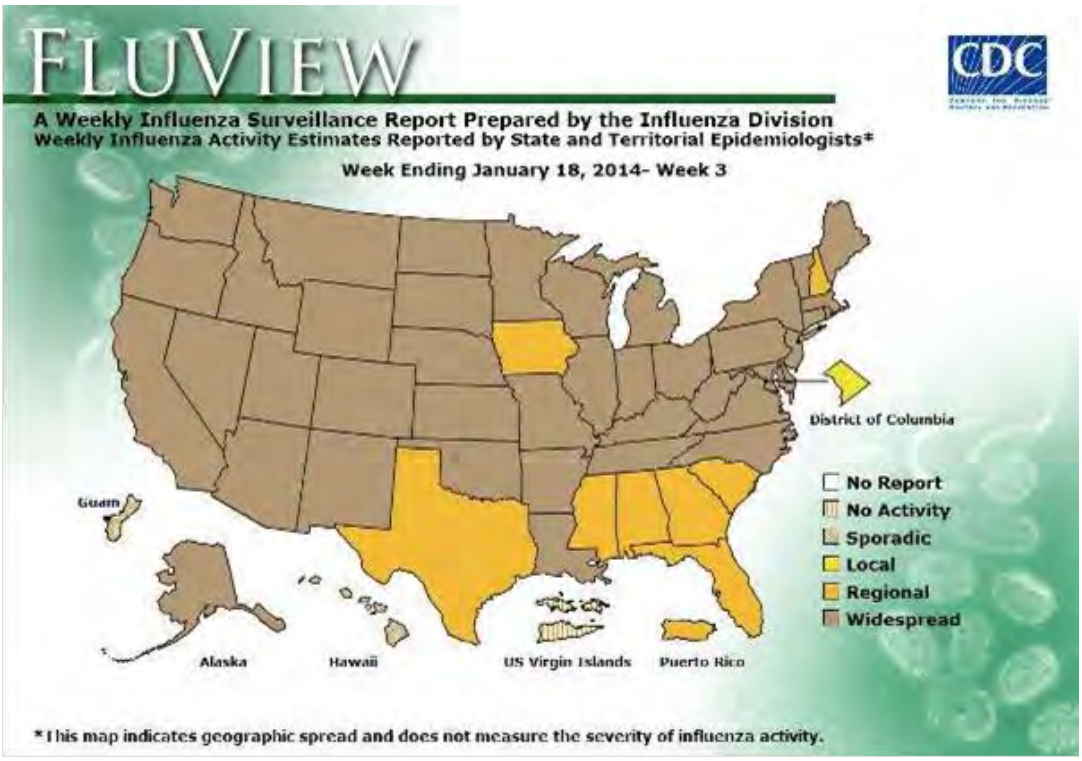

Figure 2. Geographic Spread of Influenza in the U.S. at Week Ending January 18, 2014 (from CDC 2014)

3. Traditional Surveillanc Tools (CDC)

The CDC has a collection of surveillance networks that collects and compiles various influenza-related statistics for the U.S. Various healthcare partners, ranging from clinical laboratories to clinics, have collaborated with the CDC to provide data such as the number of outpatient visits with influenza-like illnesses (ILI), and the number of samples collected for laboratory testing.

The CDC publishes statistics weekly that are tabulated and reported at the national level, the regional level, and each individual U.S. state level. At the regional level, statistics are reported based on the ten regions (Figure 3) defined by the U.S. Department of Health & Human Services.

Figure 3. HHS Regions (from U.S. Department of Health & Human Services 2006)

a. WHO and NREVSS

The World Health Organization (WHO) and The National Respiratory and Enteric Virus Surveillance System (NREVSS) collaborating laboratories are located in all 50 states including Washington, D.C., These laboratories provide the CDC with the number of respiratory specimens that are tested for influenza. These include the number of specimens collected from ILI patients, as well as the number of specimens tested positive for influenza type A and B.

b. ILINet

The U.S. Outpatient Influenza-like Illness Surveillance Network (ILINet) keeps track of the number of outpatient visits that are attributed to influenza-like illnesses (ILI). The ILINet has more than 2,900 outpatient healthcare providers nationwide.

c. FluSurv-NET

The Influenza Hospitalization Surveillance Network (FluSurv-NET) consolidates the laboratory-confirmed influenza-associated hospitalizations for 13 selected states of which ten are in the Emerging Infections Program (EIP) and three are in the Influenza Hospitalization Surveillance Project (IHSP). The EIP was established in 1995 with the objective of addressing emerging infectious disease threats. Currently, the EIP contains a network of ten state health departments, including CA, CO, CT, GA, MD, MN, NM, NY, OR, and TN. The three states in the IHSP are MI, OH, and UT.

B. TWITTER

In the U.S., there are 49 million monthly active Twitter users. These users make use of this online micro-blogging platform to broadcast information through text, images, and videos. Such sharing has allowed information to flow at a faster pace, increasing the awareness of any major events. In addition, such sharing has made available data that has been used for research in various areas to obtain patterns or novelties.

1. Using the Tweets

This section discusses some of the challenges faced when using tweets for research purposes.

a. Acronyms

The evolution of text messaging has created an informal type of English language that helps to shorten the message-typing duration. The term “ill” could mean either ill or “I’ll,” two words with entirely different meanings and usage. In addition, acronyms such as “lol” and “idk,” representing “laughing out loud” and “I don’t know,” are now commonly used in text messaging.

b. Location

Twitter users have the option of declaring the location where they are based. The declared location could be a temporary place or a false or fictitious place depending on the open-mindedness or approachability of the user. In addition, Twitter does not validate

the correctness of the location field. The user can enter California as 'Ca1i4nia' as the location. This results in the omission of influenza-related tweets from the study due to the anonymity of the user's location.

Since August 2009, Twitter users have the option to tag each tweet with their current geo-location (Stone 2009). By turning on the option, the tweet will be tagged with a global position coordinate in latitude and longitude. In a study conducted by social media analytics firm Sysomos, however, this geo-location option was only used in 0.23% of all tweets (Evans 2010).

c. Meaning of Tweet

Part of speech (POS) tagging tools has been developed to parse messages to help in predicting the meaning or feeling that the messages carried. Nonetheless, these tools are never a foolproof replacement for human annotators. Some words in the English language can be used differently. The adjective 'sick' is commonly used to declare one's ailing or ill health. It could also be used to declare one's negative feeling about certain issues happening (e.g., I am sick of his manners). In today's teenage slang, it can mean "great," as "bad" can mean "good."

C. RELATED WORK

Research has been conducted to evaluate the potential of using tweets for various applications. Some have already attempted to develop models to predict the level of influenza activity.

1. Using Twitter for Sentiment Analysis

Tweets have been used by various researchers and companies to conduct sentiment analysis. Companies such as Social Mention (www.socialmention.com) and TweetFeel use real-time tweets to evaluate negative and positive feelings associated with a search term. RAND Corporation attempted to use tweets to gauge Iranian public opinion and mood after the 2009 presidential election (Elson et al. 2012).

2. Using Twitter to Gain Situational Awareness

The use of Twitter has also been extended to gain situational awareness of major events. A recent study evaluates the potential of using tweets as an indicator for an occurrence of a snowstorm disaster (Cain 2013). It carries a similar objective to this study in providing to healthcare responders first-hand information about an emergency.

3. Predicting Influenza Activity Level

Google Flu Trends (GFT) was developed by Google to predict the U.S. ILI rates based on the frequency of the terms that are searched by Google users. In the 2007–08 flu season, GFT was lauded as its estimates were highly correlated to ILI data collected by CDC's ILINet (Ginsberg, Mohebbi et al. 2009). GFT was regarded as such a big success for predicting influenza activity that the same concept is applied for the making of Google Dengue Trends.

A recent study found that GFT has been persistently overestimating the level of flu activity over time (Butler 2013). For the 2012–2013 flu season (Figure 4), GFT's prediction almost doubles CDC's estimates. GFT's failure was suggested to be indirectly caused by the increase in public awareness from widespread media coverage of 2013's severe flu season. The broadcasting of such news may have caused the public to be more fearful and conscious about flu, thus triggering an uproar of flu-related searches to gain relevant knowledge and situational awareness of the flu threat.

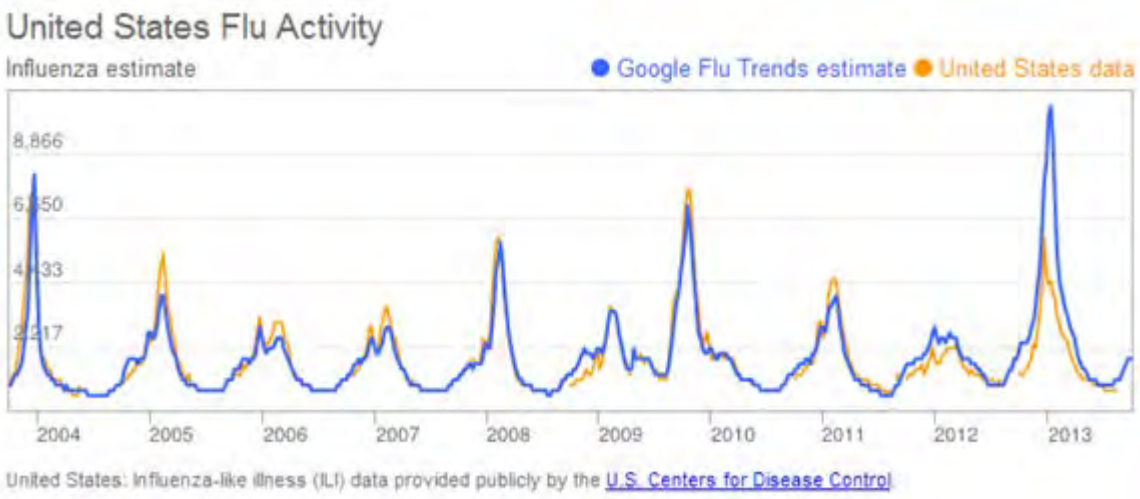

Figure 4. Historical Estimates for United States Flu Activity by Google Flu Trends (from Google 2014)

More recently, Internet search queries data of Baidu, the popular Chinese search engine, have been used to develop a model to perform a 1-month lead prediction of influenza activity in China (Yuan et al. 2013). The prediction of the fitted model is determined to be highly correlated with the surveillance data collected by China's Ministry of Health, with a mean absolute percentage error of only 10.6%.

4. Using Twitter to Predict Influenza Activity Level

A number of studies have attempted to use Twitter tweets to predict the level of influenza activity in a certain geographical area. One such study attempted to model influenza rates in the U.S. using individual keyword frequencies as predictor variables (Culotta 2010). A similar analysis conducted by Kim et al. (2013) uses tweets in the Korean alphabet, Hangul, to fit a regression model to estimate cases of influenza in South Korea. The study uses the Least Absolute Shrinkage and Selection Operator (LASSO) to select a subset of terms and uses the frequency of these terms as predictors in the regression model.

D. RESEARCH QUESTIONS

The use of manual, traditional, influenza surveillance tools, such as ILINet, typically has a 1–2 week reporting lag. Past related work, such as GFT and Culotta (2010) have attempted to construct models using Twitter messages to predict the level of influenza activity at a national level. However, to be able respond promptly and accurately, the CDC needs a tool that can predict or forecast influenza outbreaks within a smaller geographical region.

Many important insights and solutions to problems have been inferred from research conducted through Big Data analytics. Is Big Data analytics a credible solution for the prediction of influenza activity in the U.S.? With the large amount of U.S. Twitter users (6th highest in terms of percentage of population), can the tweets be used to predict the level of influenza activity? What kind of estimates can it provide to the CDC?

This study attempts to evaluate the feasibility of using tweets as a leading indicator of influenza activities in the U.S. Models are developed using regression analysis to predict the level of influenza activity at a national , regional and state level.

The tweets are first transformed into statistics representing the predictor variables. Influenza-related statistics are collected from CDC to represent the response variables. The predictor variables and response variables are then populated into a weekly time series dataset. The dataset is further divided into a training set and a test set. The best prediction models are then constructed using the training set with the best subset of predictors identified via variable selection techniques. The independent test set is then used to validate the prediction models.

Using the test set predictions obtained from the best models, the correlation between the statistics drawn from the tweets and the actual CDC surveillance data are then evaluated. The presence of correlation would support the idea of using tweets as a leading indicator of influenza activity level in the U.S.

THIS PAGE INTENTIONALLY LEFT BLANK

III. DATA

This chapter describes the data collection method for the two types of data collected for this study, namely the tweets and the influenza-related statistics.

A. TWEETS

From September 2012 to June 2014, a database of 22 months of tweets was collected and provided by the Santa Fe Institute. Within the database, however, there are several months of missing tweets. Figure 5 shows the number of days with missing tweets for each month. It is unfortunate that the periods of missing tweets occur during the month of January (in 2013) and February (in 2014) where influenza activities typically peak.

Figure 5. Number of Days with Missing Tweets

Each tweet is stored as a JavaScript Object Notation (JSON) object in the data. JSON is a language-independent text format that is widely used for transferring data. JSON objects can be created and parsed by many programming languages, including R.

The R “rjson” package contains the fromJSON function that converts a JSON object into its corresponding R object (Couture-Beil 2014). The created R object exists in the form of a vector where each element is accessible via an index. Three elements are necessary for this study: the text message, the user specified location, and the time zone.

B. INFLUENZA ACTIVITY LEVELS

The CDC monitors the health status of the U.S. public and provides information such as emergency preparedness, disease-related statistics, and facts relevant to various types of diseases. Seasonal influenza is one such disease that CDC is monitoring with caution. It publishes weekly reports on flu activity in the U.S. These reports provide statistics such as the number of ILI outpatient visits and number of positively tested samples. In addition, archives of these statistics are tabulated and made publicly accessible in the form of downloadable Microsoft Excel files.

1. ILI Outpatient Visits

ILI outpatient visit statistic is compiled with the help of healthcare providers (clinics) who report any ILI patient visits to the ILINet. From this reporting, data such as the number of ILI outpatient visits, as well as the percentage of ILI visits, are made available. Figure 6 shows a chart provided on CDC’s website; it plots the weekly percentage of ILI visits for the 2013–2014 influenza season.

Figure 6. Percentage of ILI Outpatient Visits for 2013–14 Flu Season (from CDC 2014)

2. Respiratory Specimens Collected and Tested Positive for Influenza Type A or B

As collaborating laboratories, WHO and NREVSS report to the CDC the number of respiratory specimens that are tested for influenza and the positive numbers for each type and subtype of influenza. Figure 7 shows a chart provided on the CDC's website; it plots the percentage of samples that tested positive as well as the samples that tested positive for various influenza types during the 2013–2014 influenza season.

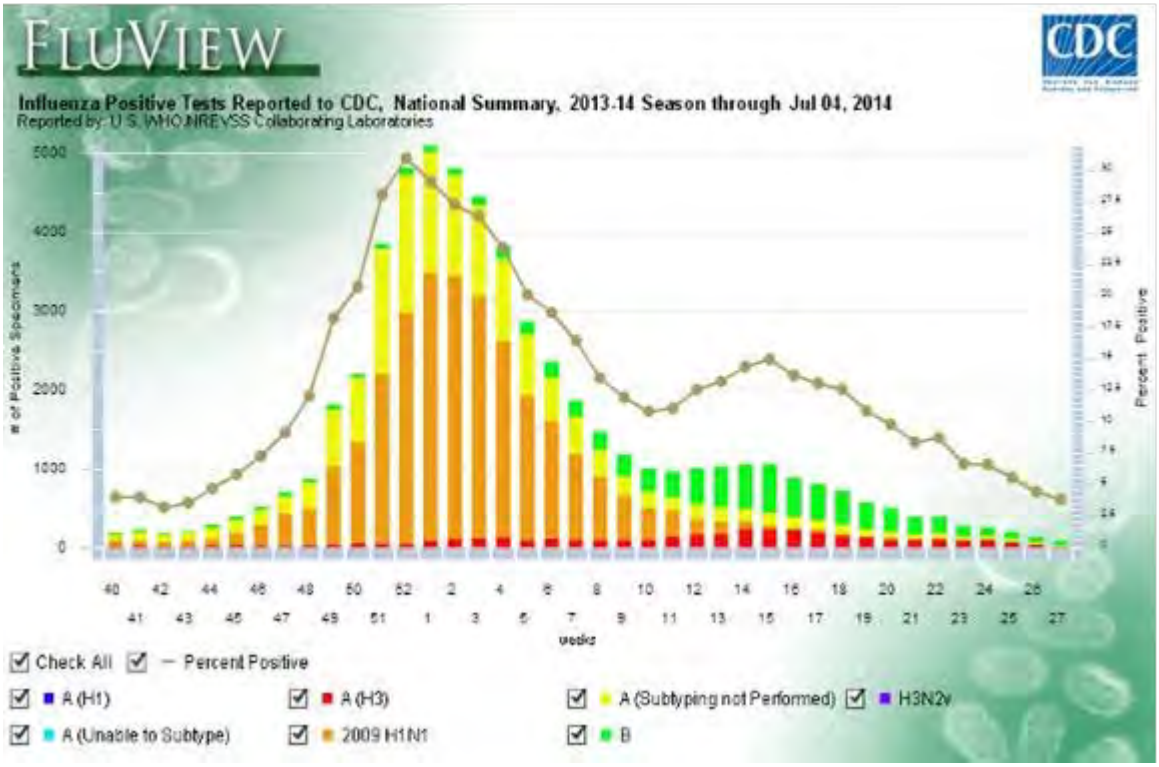

Figure 7. Influenza Positive Tests for Respiratory Specimens Collected for 2013–14 Flu Season (from CDC 2014)

3. Influenza Associated Hospitalizations

During each flu season, the Influenza Hospitalization Network (FluSurv-NET) monitors the influenza-associated hospitalizations in the U.S. over a period of 29 weeks (starting from early October). Hospitalizations that are laboratory confirmed to be influenza-associated are reported to FluSurv-NET.

Figure 8 shows the weekly rate of influenza-associated hospitalizations per 100,000 population for the past four flu seasons as published by the CDC. It also shows that the sharp increase in the number of influenza-associated hospitalizations coincides with the peak (January) of the flu season.

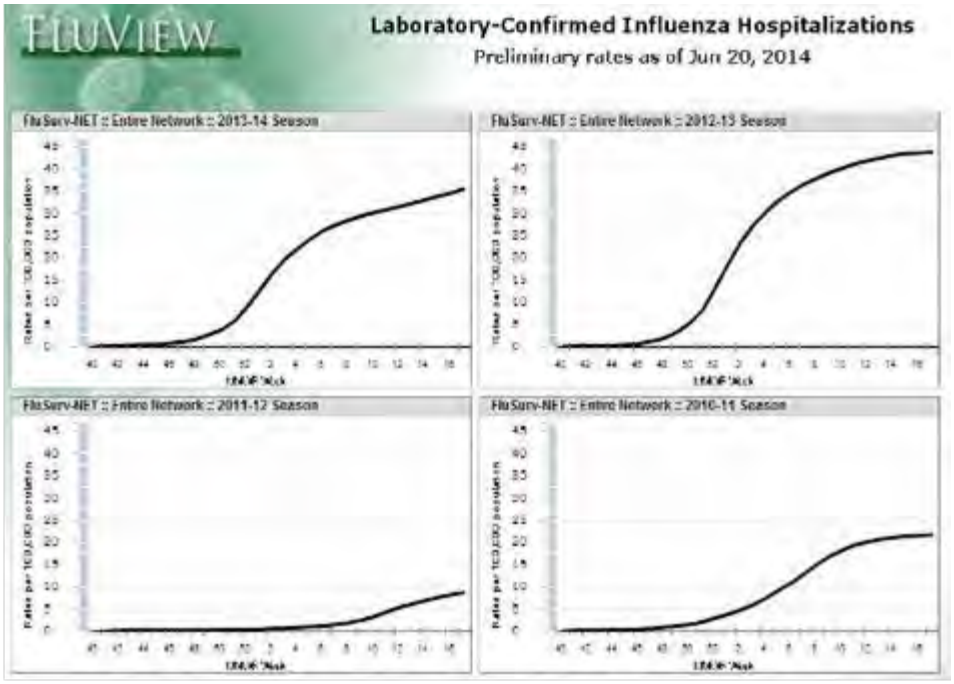

Figure 8. Rate of Laboratory-Confirmed Influenza Hospitalizations for 2013–2014 Season (from CDC 2014)

THIS PAGE INTENTIONALLY LEFT BLANK

IV. APPROACH

A model is fitted using a regression analysis to estimate the following three types of response variables: number of influenza-associated hospitalizations, number of ILI outpatient visits, and number of positively tested samples. As for the predictor variables, the study explores the correlation of counts obtained for various categories of words, such as flu symptoms, flu complications, and flu-related activities against the three types of response variables. The following sub-sections shall elaborate on the approach of the study.

A. REGRESSION

Regression is an approach for modeling relationships between a numeric response and predictor variables. Given a collection of observed data, regression fits a model in the form of an equation that can be used to predict the value of the response variable given the values of the predictor variables.

In this approach, the observed data is a sequence of data points accumulated for each week. It is also known as a time series dataset. The individual weekly counts for each category of words are accumulated and stored in the time series dataset. The time series dataset also stores the response variables that are extracted from CDC Influenza Surveillance networks.

Regression analysis is then performed using the weekly time series dataset to find a best subset of predictor variables to fit the equation. Time series analysis is an alternative technique that can be used to construct models by lagging the predictor variables. In this study, however, each data point is assumed to be independent and identically distributed. In other words, the count of a predictor variable for a specific week is not related to the past or subsequent count of the same predictor variable.

In this study, all required preprocessing of data and regression analysis are performed using R. R is a language and environment that provides statistical and graphical techniques suitable for data analysis (R Core Team 2013). The final R source

code that implements the approach of this study can be obtained from Samuel Buttrey (Associate Professor of Operations Research at the Naval Postgraduate School).

1. Predictor Variables

The predictor variables are categorized into three types: indicative, supportive and general. The indicative and supportive predictor variables are frequencies of categories of keywords. The general predictor variables accumulate the count of tweets that are influenza-related.

Past research has used the frequencies of individual keywords as predictor variables. In a study conducted in Korea, Kim et al. (2013) begin with an initial set of 500 Hangul (Korean alphabet) keywords that were chosen based on their high frequencies of appearing in influenza-related tweets. The study manages to obtain an R^2 of 0.998 by fitting a model with the frequencies of 60 keywords selected via LASSO.

In contrast, this study attempts to use frequencies of categories of hand-chosen terms as predictor variables. This approach aims to "reward" tweets that contain sentences phrased using the basic sentence unit. The presence of a term "flu" or "influenza" in a sentence might mean the occurrence of a flu-related event; it might not mean a flu attack on a person, however. If the sentence is phrased using a combination of pronouns, adjectives, and verbs that are typically used by someone to express their ill-being, we can be more certain that the sentence indicates a flu attack.

English words are categorized into eight categories (also known as parts of speech): noun, pronoun, adjective, verb, adverb, preposition, conjunction, and interjection. In a basic sentence unit, there is typically a subject, an action verb, and an object. Both the subject and the object are typically nouns that represent a person, a place, or a thing. In addition to nouns, adjectives may be included in a sentence to describe the person's feeling.

a. Indicative Predictors

In the context of this study, nouns and adjectives that are indicative of a flu attack in a sentence are grouped into five categories. (I1) Flu.Activities consist of words relating

to activities carried out by the influenza-affected patient. (I2) Flu.Terms consist of common terms of influenza such as flu, H1N1, and viral infection. (I3) Flu.Symptoms consist of words that are symptoms of influenza such as fever and body aches. (I4) Medicines lists common medicines that are prescribed to or purchased by a flu patient. (I5) Flu.Complications gives common terms regarding flu complications. Flu-related complications include pneumonia, bronchitis, sinus infections, and ear infections. Children younger than five years and adults older than 65 years are at a higher risk of suffering from these complications. The list of keywords selected for each category can be found in Appendix B.

In summary, the counts of the five categories of keywords are the five "Indicative" predictor variables. For this study, each tweet that has at least one "Indicative" term(s) is considered as an influenza-related tweet.

b. Supportive Predictors

The next five categories of terms are terms that cannot indicate occurrences of flu attack independently. These terms, however, can certainly support the claim of an influenza-related tweet that already has one or more "Indicative" terms.

(S1) Verbs consists of verbs that are used to convey actions related to a flu attack. (S2) Pronouns consists of pronouns that are commonly used in place of nouns as the subject in a basic sentence unit. The existence of a pronoun might support the assumption that a person has been involved in a flu attack. (S3) Adjectives consists of adjectives that a flu-affected patient would use to describe his or her ill-being such as unwell, severe or worse. (S4) Emoticons is included with the aim of "rewarding" the use of emoticons to express one's feelings or mood. In text messaging, the absence of one's body language or facial expression is typically replaced by both adjectives and emoticons. Hence, the presence of emoticons that express sadness (Table 1) in an influenza-related tweet supports the assumption that a person is troubled by flu illnesses. (S5) Rest.Activities is included to count the occurrences of rest days taken by a flu-hit patient. These rest days could come in the form of employer-granted days off or doctor-issued medical

certificates. In summary, the counts of the five categories of keywords are the five "Supportive" predictor variables.

Emoticons (Rotated 90^{o}) Expressing Sadness				
:-[	:-(	:(	:-c	:[
:c	:-<	:-[	:<	:-{

Table 1. Emotions that Expressed Sadness (from Wikipedia Contributors 2014)

c. General Predictors

The final eight predictors are used to accumulate the number of tweets in eight different ways. The predictor Influenza.Related.Tweets accumulates the number of influenza-related tweets. An influenza-related tweet is previously defined as a tweet that has at least one "Indicative" term(s). The next seven predictors of this category comprise the number of influenza-related tweets of a particular number of matching term(s). The first six predictors are introduced and named as X*.Term(s) where the * denotes the number of matching term(s). Influenza-related tweets that contain seven or more matching terms are accumulated using the predictor X7.Terms.

2. Keyword Selection

No.	Predictor	Type	Descriptions
1	Flu.Activities	Indicative	Terms related to activities carried out by an influenza-hit patient. E.g., hospital or clinic visit
2	Flu.Terms		Terms related directly to influenza
3	Flu.Symptoms		Terms related to symptoms of influenza
4	Medicines		Terms related to medicine typically issued for patients with flu-related illness
5	Flu.Complications		Terms related to flu complications
6	Verbs	Supportive	Verbs: Action words that are used to describe the above five categories
7	Adjectives		Adjectives: Descriptive words used to describe a person's well being

No.	Predictor	Type	Descriptions
8	Pronouns		Terms defining the subject or object in a clause
9	Emoticons		Emoticons used to describe sender's sadness in a message
10	Rest.Activities		Terms related to rest days afforded to patient
11	Influenza.Related.Tweets	General	Total number of influenza-related tweets
12	X1.Term		A total of seven predictor variables that count the total number of influenza-related tweets. Each predictor variable accumulates the total for a different number of term(s) matching categories 1–10.
13	X2.Terms		
14	X3.Terms		
15	X4.Terms		
16	X5.Terms		
17	X6.Terms		
18	X7.Terms		

Table 2. List of Predictors for Regression Analysis

Table 2 shows the final list of 18 predictor variables selected for regression analysis. Hand-chosen terms for the "Indicative" and "Supportive" categories are carefully selected from two sources: healthcare services websites and the tweets. Healthcare services websites such as the CDC are also browsed to obtain terms such as the list of flu symptoms and flu complications, and the list of medicines or remedies for flu.

Figure 9. Results Obtained for a Keyword Search: "Down with Flu"

Real-time tweets are randomly searched and sampled for terms that indicate a person having influenza-like illnesses. Figure 9 shows a subset of results returned from a key phrase “Down with flu” search using the Twitter search function available to all Twitter account holders. All three tweets indicate occurrences of flu-related activity. Table 3 shows an illustration of choosing terms for the ten categories using the second tweet as shown on Figure 9. Refer to Appendix C to view the complete list of terms chosen for each of the ten categories.

Terms in Tweet	Category	Terms in Tweet	Category
Down	COUNT_VERB	it	COUNT_PRONOUN
with	None	from	None
flu	COUNT_FT	Kemal	None
again	None	this	None
I	COUNT_PRONOUN	time	None
think	None	--”:(	COUNT_EMOTIONS
got	COUNT_VERB		

Table 3. Illustration of Selecting Terms to Match for the Ten Categories

3. Definition of an Influenza-Related Tweet

This approach considers a tweet to be influenza-related only if it has at least one “Indicative” term(s). The terms in a tweet are first matched against the terms in the Indicative categories. The presence of “Indicative” terms will then subject the terms in the tweet to matching against the terms in the Supportive categories. Otherwise, the tweet will be regarded as an irrelevant tweet and omitted from the study.

4. Weekly Time Series Dataset

Each individual tweet undergoes the matching and counting process against the ten categories of terms. These counts obtained from the influenza-related tweets are then accumulated as a weekly time-series dataset. Each week is deliberately set to start from a Sunday and ends on the following Saturday to correspond to the CDC’s weekly standard of reporting and archiving flu statistical data.

There are three separate datasets for accumulating counts at the national level, HHS regional level, and state level. The location (if provided by the tweet originator) and time zone that is tagged to the relevant tweet determines the dataset(s) onto which to accumulate the counts contributed by the tweet.

The time zone of a tweet posted in the U.S. is tagged with the name of one of the six time zones followed by "(U.S. and Canada)," e.g., "Pacific Time (U.S. and Canada)." Tweets that are tagged with any U.S. time zone contribute to the dataset for the national level. As Canada also uses U.S. time zones, a certain percentage of the relevant tweets may have originated from Canada. This percentage, however, may prove to be insignificant due to Canada's relatively small percentage of Twitter users compared to the U.S. (Beevolve 2012).

The location specified by the tweet originator is tagged to the tweets that they posted. The specified location is matched against the names of the states (or their abbreviations, e.g., California → CA) in the U.S. or the names of the top 50 most populated cities in U.S. If there is a matching city or state for the location of a tweet, the count of the tweet will be updated to the dataset for the state level and regional level. Because Twitter does not validate the entry for the location field, tweets with unspecified, fictitious, or inaccurately spelled location entries are omitted from the study.

5. Response Variables

This study selected the following four types of response variables (Table 4) for regression analysis.

Type	Response Variable	National Level	HHS Regional Level	State Level
1	Number of outpatient ILI visits	✓	✓	
3	Number of collected respiratory specimens	✓	✓	
2	Number of respiratory specimens tested positive for influenza type A or B	✓	✓	
4	Number of influenza-associated hospitalizations			✓

Table 4. Response Variables for Regression Analysis

6. Fitting Models

The weekly time series datasets are first tabulated with the response and predictor variables that are collected from the CDC and counted from the tweets. Next, subsets of the datasets are created to be used as a training set and a test set. Using the training set, prediction models are then constructed with the best subset of predictor variables identified through the exhaustive search variable selection technique as well as cross-validation. These prediction models are then further validated using the test set.

a. Training Set

A subset of the time series dataset is used for training and fitting the prediction model. For the models constructed to predict the number of outpatient ILI visits and respiratory specimens (response variables), the training set is allocated as 51 weeks of tweets ranging from February 2, 2012 to January 24, 2014. The corresponding values of the response variables for the 51 weeks are collected from the CDC's website and included into the dataset. Five weeks of data were excluded, however, from the training set due to the unavailability of tweets. The training set for constructing models to predict the number of influenza-associated hospitalizations is allocated as 30 weeks of tweets ranging from October 5, 2012 to April 19, 2013.

b. Test Set

A subset of the time series datasets is allocated as a test set. The test set is used to assess the prediction model constructed using the training set. As the test set should be

independent of the training set, it is allocated as 18 weeks of tweets ranging from February 22, 2014 to June 27, 2014. The corresponding values of the responses for the 18 weeks are collected from the CDC's website. Two weeks of data, however, were excluded from the test set due to the unavailability of tweets.

c. Variable Selection—Exhaustive Search

Variable selection is a process where the best subset of predictor variables is selected. This process helps in the selection of the most influential predictor variables from a huge set, as well as the omission of redundant or excessive predictor variables that cause collinearity. Testing-based procedures such as backward elimination, forward selection, and stepwise regression are procedures that are performed manually. The statistician decides which variable to eliminate or select during each iteration. This manual process stops when the statistician feels that they have arrived at the best set of predictors.

The R package leaps has a function regsubsets (regression subsets) that returns the best subset based on exhaustive search, forward selection, or backward elimination (Lumley, Thomas using Fortran code by Miller, Alan 2009). This package aids in automating the variable selection process. This study uses the regsubsets function to return the best subsets for sizes ranging from one to eight predictor variables based on exhaustive search. The exhaustive search is a brute force search that builds every possible regression model from the given set of predictors and recommends the subset of predictors that has the highest coefficient of determination (R^2). Hence, the exhaustive search is considered a more comprehensive testing-based procedure than forward selection and backward elimination.

d. Cross-Validation

Cross-validation is a regression model validation technique. In general, the technique uses a subset of a dataset to fit a model. This model is then validated using the remaining portion of the dataset (commonly known as validation set).

In this study, the training set, defined in IV.A.6.a, is subjected to a ten-fold cross-validation. The training set is randomly partitioned into ten equal-sized subsets. Each subset is then used as a validation set once for testing against the model constructed using the other nine subsets. This technique is repeated 100 times with each set of repetitions averaged to give the average standard deviation of residuals.

A homegrown R function BestsubXval (Koyak 2013) is developed for use in conjunction with the output of the regsubsets function. The function takes in the list of best subset of predictors returned by regsubsets function, performs a ten-fold cross-validation for the list using another homegrown R function xval (Buttrey 2012) and returns the average standard deviation of residuals for each subset of predictors. The best subset of predictors is then identified as the subset with the smallest average standard deviation of residuals.

V. ANALYSIS

This section presents the results and discusses the findings for the prediction models that are constructed for each of the four response variables. The section is further divided into four sub-sections. Each sub-section discusses the models that are constructed for a response variable at the national level, regional level, and/or state level.

Each constructed model is evaluated by looking at its adjusted R^2, its p-Value as well as analysis of residuals (errors). In this study, as the total number of weekly observations is about 70 (which is considered small), the adjusted R^2 (adj.R^2) is used as the measure to evaluate the fit of each constructed model instead of the coefficient of determination (R^2). A constructed model is typically considered to be well fit if it achieves an adj.R^2 of at least 0.6. Each model is further refined after performing residual and outlier analysis.

The Pearson's correlation coefficient is used as the primary measure for determining the correlation between the predicted values and actual CDC influenza surveillance data. The predicted value for each week in the training and test set are compared against the actual value for the same week to derive the Pearson's correlation coefficient.

A. MODEL FOR PREDICTING NUMBER OF OUTPATIENT ILI VISITS

The resulting model for the national level (NL) seems to suggest the presence of correlation between the Twitter messages and the influenza activity level. The NL model gives a fairly good prediction, capturing the increasing and declining trend of the number of outpatient ILI visits throughout the U.S. flu season.

The Pearson's correlation coefficient between the test set predictions of the NL model and actual CDC ILI surveillance data is computed to be 0.900 (95% CI: 0.732, 0.965). Furthermore, after combining the training and test set predictions, the Pearson's correlation coefficient between the combined set and the actual CDC ILI surveillance data is computed to be 0.905 (95% CI: 0.846, 0.942). Unlike the NL model, however, the

constructed models for each of the ten individual HHS regions have varying results. The models appear to be well fit only for four regions.

1. Model for National Level

Models are constructed to predict the number of outpatient ILI visits using the best subset of predictors that is identified through the variable selection process described in IV.A.6. Figure 10 and Figure 11 show the relationship between each Indicative and Supportive predictor variable against the number of outpatient ILI visits. From the scatterplots, Flu.Terms and Flu.Complications seem to be the only predictors that are clearly correlated to the number of visits. The trend lines (in blue), however, do indicate that each predictor variable is positively correlated to the number of visits.

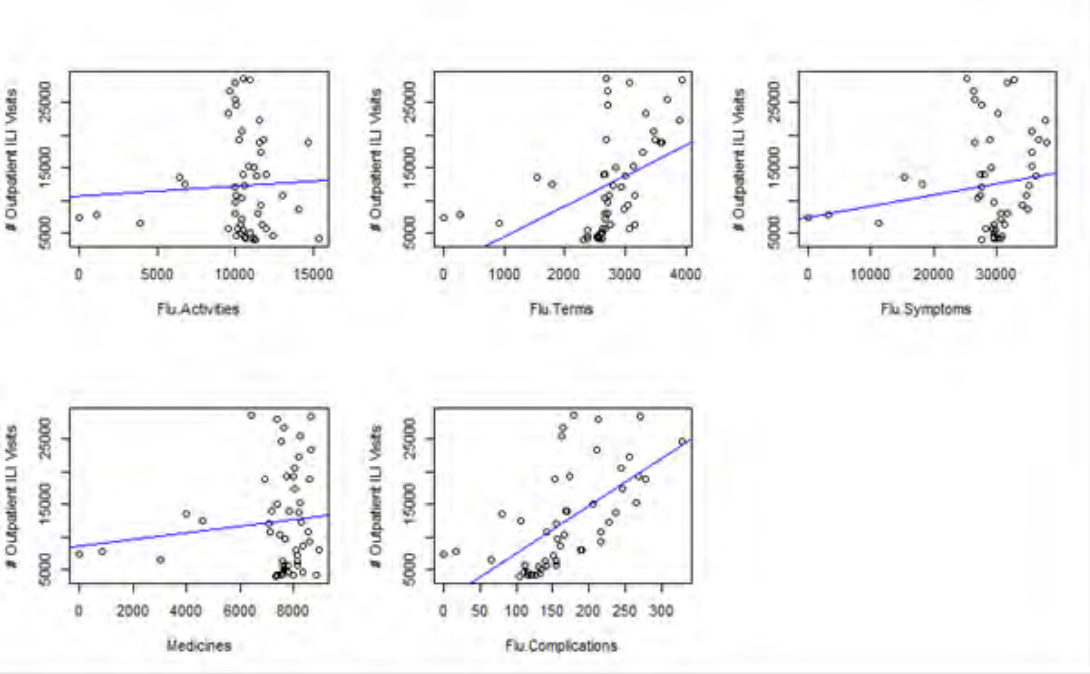

Figure 10. Relationship between Each Indicative Predictor Variable and the Number of ILI Outpatient Visits

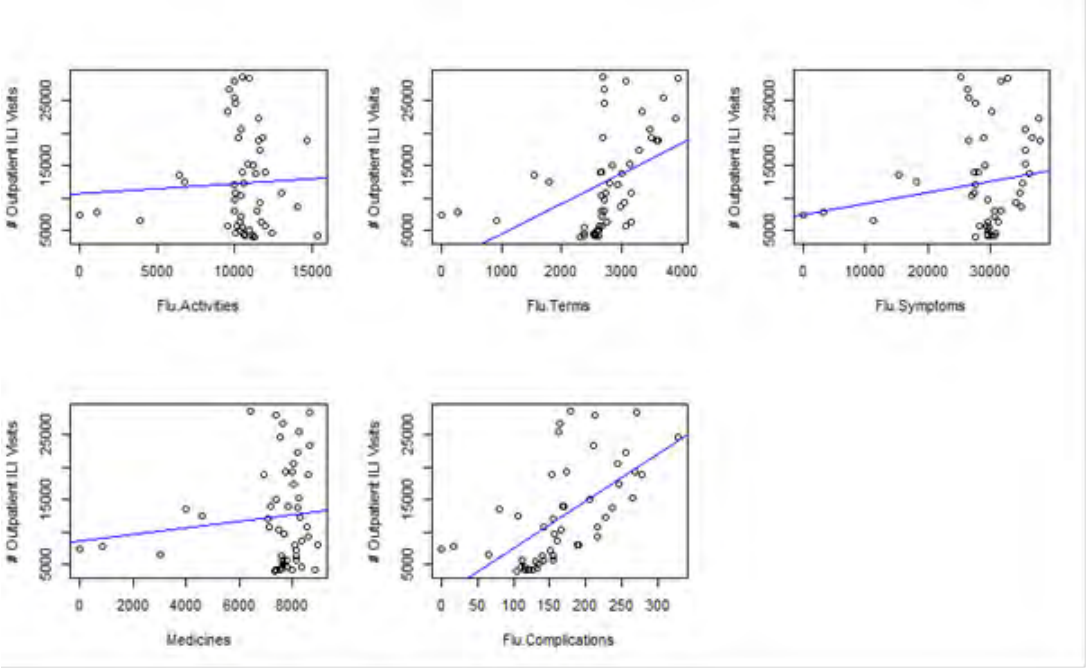

Figure 11. Relationship between Each Supportive Predictor Variable and the Number of Outpatient ILI Visits

Table 5 shows the eight best subsets of predictors that are returned by the exhaustive search algorithm. The best subset among the eight is then obtained after performing a ten-fold cross validation. In this case, the 7th subset (denoted with *) is identified as the best subset with the smallest average standard deviation of the residuals of 4159. Figure 12 shows the statistical summary of the model that is constructed using the 7th subset.

Predictors	Subsets of Predictors							
	1	2	3	4	5	6	7*	8
Flu.Terms			*	*	*	*	*	*
Verbs						*	*	*
Adjectives							*	*
Pronouns		*	*					
Flu.Complications	*	*	*	*	*	*	*	*
Emoticons					*	*	*	*
Influenza.Related.Tweets						*	*	*
1.Terms				*	*			
6.Terms								*
7.Terms				*	*	*	*	*
Average Residual Standard Error	5971	5350	4392	4304	4507	4255	4159	4382

Table 5. Best Subsets of Predictor Variables (Original)

```
Residuals:
    Min      1Q  Median      3Q     Max
-5807.4 -1963.9    69.1  1560.8  9256.8

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              10534.6889  8342.3965   1.263 0.214359
Flu.Terms                    7.4364     1.7923   4.149 0.000181 ***
Verbs                        4.1574     1.5096   2.754 0.008984 **
Adjectives                   8.4627     5.4157   1.563 0.126429
Flu.Complications           81.5521    14.6816   5.555 2.32e-06 ***
Emoticons                   13.7369     4.6213   2.973 0.005104 **
Influenza.Related.Tweets    -2.0023     0.4259  -4.701 3.36e-05 ***
X7.Terms                   -47.0675    12.3865  -3.800 0.000509 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3352 on 38 degrees of freedom
Multiple R-squared:  0.8489, Adjusted R-squared:  0.8211
F-statistic: 30.51 on 7 and 38 DF,  p-value: 1.016e-13
```

Figure 12. Statistical Summary of Constructed Model for Number of Outpatient ILI Visits

The model constructed using the best subset achieves a high adj.R^2 of 0.8211, which indicates that that the model is well fit. In addition, the model has a small p-Value of 1.016e-13 that corresponds to a high level of confidence in terms of making predictions. Figure 13 shows the equation for predicting the number of outpatient ILI patients.

$$\begin{aligned}\#\text{Outpatient.ILI.Patients} = {} & 10534.69 + 7.44 * \text{Flu.Terms} + 4.16 * \text{Verbs} + 8.46 * \text{Adjectives} \\ & + 81.55 * \text{Flu.Complications} + 13.74 * \text{Emoticons} \\ & - 2.00 * \text{Influenza.Related.Tweets} - 47.07 * \text{X7.Terms}\end{aligned}$$

Figure 13. Equation for Predicting Number of Outpatient ILI Patients

The most significant and influential predictors are Flu.Terms and Flu.Complications. This is not surprising given the fact that the words in both categories are terms that are directly related to influenza. In addition, both Verbs and Emoticons also appear to be statistically significant in their contribution to the number of outpatient visits.

Another interesting observation is the negative correlation between the number of Influenza.Related.Tweets and X7.Terms against response. It is observed from the dataset (from February 2013 to June 2014) that the number of influenza-related tweets and tweets with seven or more matching terms has massively decreased by two- to four-fold from the beginning of February 2013 to June 2014. This may indicate either a decrease in users' participation in Twitter or the unintended inclusion of selected keywords that are commonly used in non-influenza-related events.

Figure 14 shows the predicted and actual number of outpatient ILI-related visits for the training set and test set. The predictions made for the training set did decline and increase as expected. Generally speaking, they did capture the trend of the influenza season. As for the test set, it is limited to just 16 weeks, which coincides with the decline in influenza activity level. Still, its predictions are encouraging as they did match a declining trend.

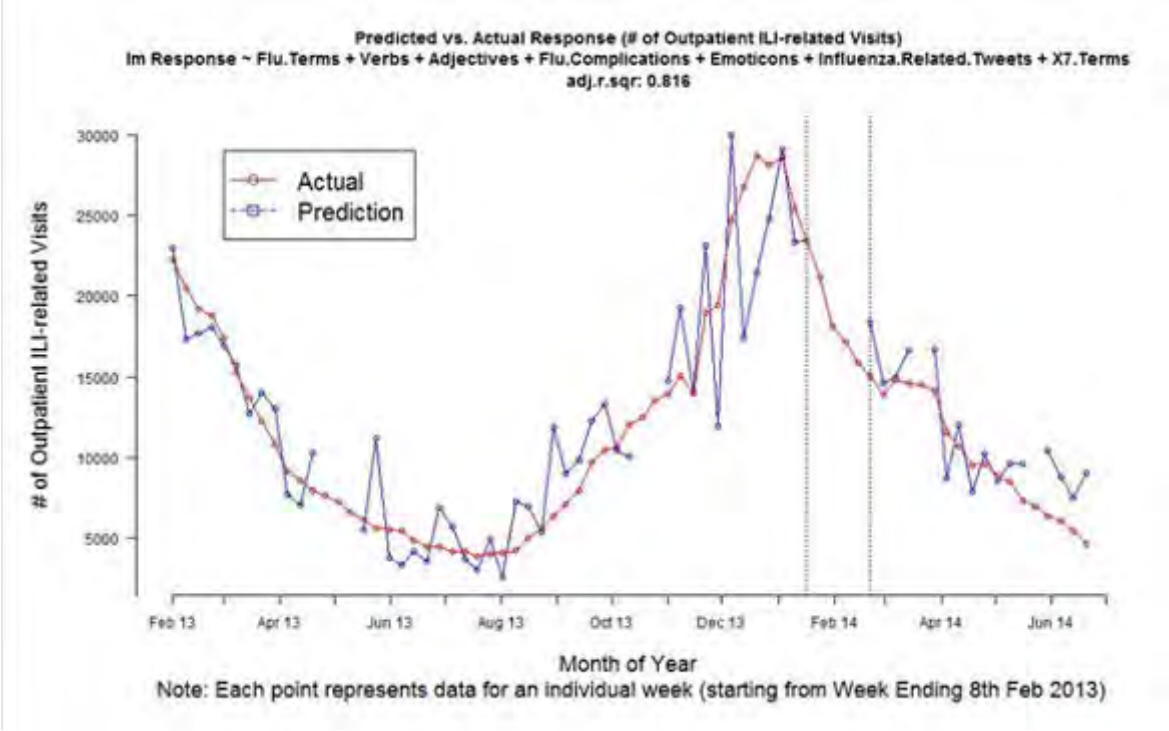

Figure 14. Predicted vs. Actual Values for Number of Outpatient ILI Visits

While the model is able to exhibit high correlation of the predictions to the actual outpatient ILI visits, the residuals (difference between the actual and predicted value) still seem to be too high with a standard deviation of 3352, which is relatively large. At the national level, the error of underestimating (or overestimating) as seen from the large residuals may not be significant due to the size of the U.S. population. At the regional level, however, it will not be practical to use a model that predicts with such a wide range of error.

The plot of residuals versus predicted values in Figure 15 shows the size of the residual for each data point (week). There are a couple of underestimated predictions occurring in the month of December 2013. The underestimated predictions are further examined and found to be caused by a sharp variation in usage of terms in at least one of the following predictors: Flu.Terms, Flu.Complications and Emoticons.

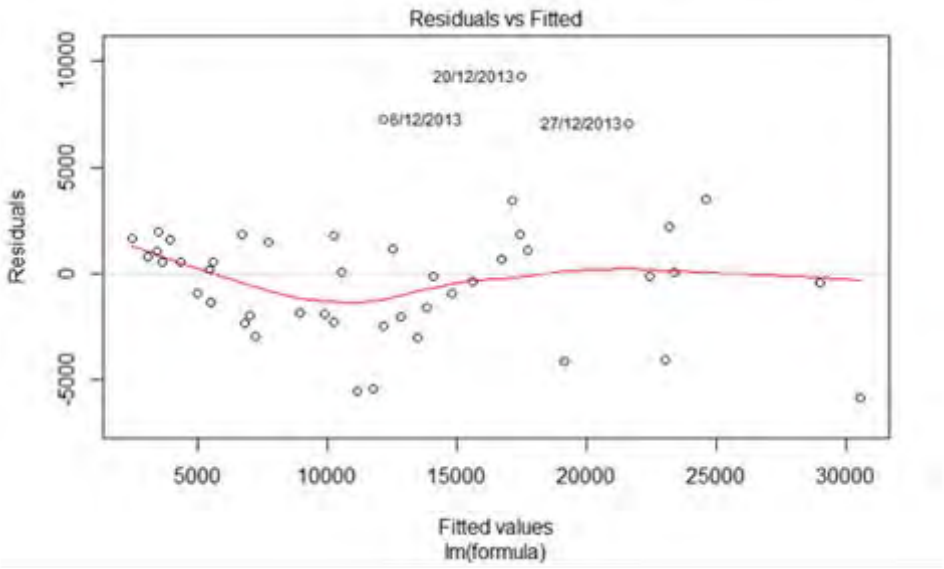

Figure 15. Residuals vs. Fitted (Predicted) Values for Constructed Model

2. Refined Model for National Level

A refined national model is constructed with the exclusion of one outlier (data point on week ending 20/12/2013). The model achieves an adj.R^2 of 0.852 as compared to 0.821 of the original model. In addition, an improvement is also seen through its Pearson's correlation coefficient of 0.900 (95% CI: 0.732, 0.965) as compared to the original model's 0.805 (95% CI: 0.516, 0.930). Table 6 shows the eight best subsets of predictors that are returned by the exhaustive search algorithm.

Predictors	Subsets of Predictors							
	1	2	3	4	5	6	7*	8
Flu.Terms			*	*	*	*	*	*
Medicines								*
Verbs					*	*	*	*
Adjectives							*	*
Pronouns		*	*	*	*			
Flu.Complications	*	*	*	*	*	*	*	*
Emoticons				*	*	*	*	*
Influenza.Related.Tweets						*	*	
2.Terms								*
7.Terms						*	*	*
Average Residual Standard Error	5541	5077	3917	3947	3763	3815	3763	3788

Table 6. Best Subsets of Predictor Variables (Refined)

The same process of performing variable selection and the ten-fold cross validation returns the 5th and 7th subset (denoted with * in Table 6) as the best subsets, with both having the smallest average standard deviation of residuals of 3763. Hence, models for both subsets are constructed to analyze and determine the best among the two. The Pearson's correlation coefficient between the test set predictions generated from each model and the actual CDC ILI surveillance data are computed to be 0.896 (95% CI: 0.719, 0.930) for the 5th subset and 0.900 (95% CI: 0.0.732, 0.965) for the 7th subset, respectively. Thus, based on the Pearson's correlation coefficient, the model constructed using 7th subset is slightly better than the 5th. Figure 16 shows the statistical summary of the refined model.

```
Residuals:
    Min      1Q  Median      3Q     Max
-5155.2 -1831.1   397.8   821.0  8580.5

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             3513.6738  7693.8266   0.457 0.650566
Flu.Terms                  7.2687     1.5914   4.567 5.32e-05 ***
Verbs                      3.2898     1.3645   2.411 0.020989 *
Adjectives                 7.8410     4.8098   1.630 0.111538
Flu.Complications         83.0680    13.0372   6.372 1.97e-07 ***
Emoticons                 15.3140     4.1281   3.710 0.000678 ***
Influenza.Related.Tweets  -1.8242     0.3817  -4.779 2.78e-05 ***
X7.Terms                 -38.6416    11.2760  -3.427 0.001511 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2975 on 37 degrees of freedom
Multiple R-squared:  0.8751, Adjusted R-squared:  0.8515
F-statistic: 37.04 on 7 and 37 DF,  p-value: 7.807e-15
```

Figure 16. Statistical Summary of Constructed Model (Refined) for Number of Outpatient ILI Visits

The standard deviation of the refined model is 2975, which is smaller than 3352 from the original model. This improvement also indicates the improved fit of this model. Figure 17 shows the prediction equation that is derived from the model constructed using the 7th subset of predictors.

$$\begin{aligned}\#\text{Outpatient ILI Visits} = 3513.67 &+ 7.27 \times \text{Flu.Terms} + 3.29 \times \text{Verbs} + 7.84 \times \text{Adjectives} \\ &+ 83.07 \times \text{Flu.Complications} + 15.31 \times \text{Emoticons} \\ &- 1.82 \times \text{Influenza.Related.Tweets} - 38.64 \times \text{X7.Terms}\end{aligned}$$

Figure 17. Equation (Refined) for Predicting Number of Outpatient ILI Patients

Additional iterations of excluding outliers are also carried out after seeing improvements from the outlier exclusion. In general, the fit of the model improves significantly with the exclusion of more outliers, but deteriorates in its precision of prediction when validated using the test set.

The exclusion of more outliers does not correspond to a lower Pearson's correlation coefficient between the overall predicted values and actual CDC ILI

surveillance data. Hence, we deduce that further exclusion of outliers will only improve the fit but not the correlation coefficient.

Figure 18 shows the actual CDC ILI Surveillance data and the predicted values generated from the original and refined models. The refined model seems to be able to predict with better accuracy for most of the weeks, with the exception of the last two months.

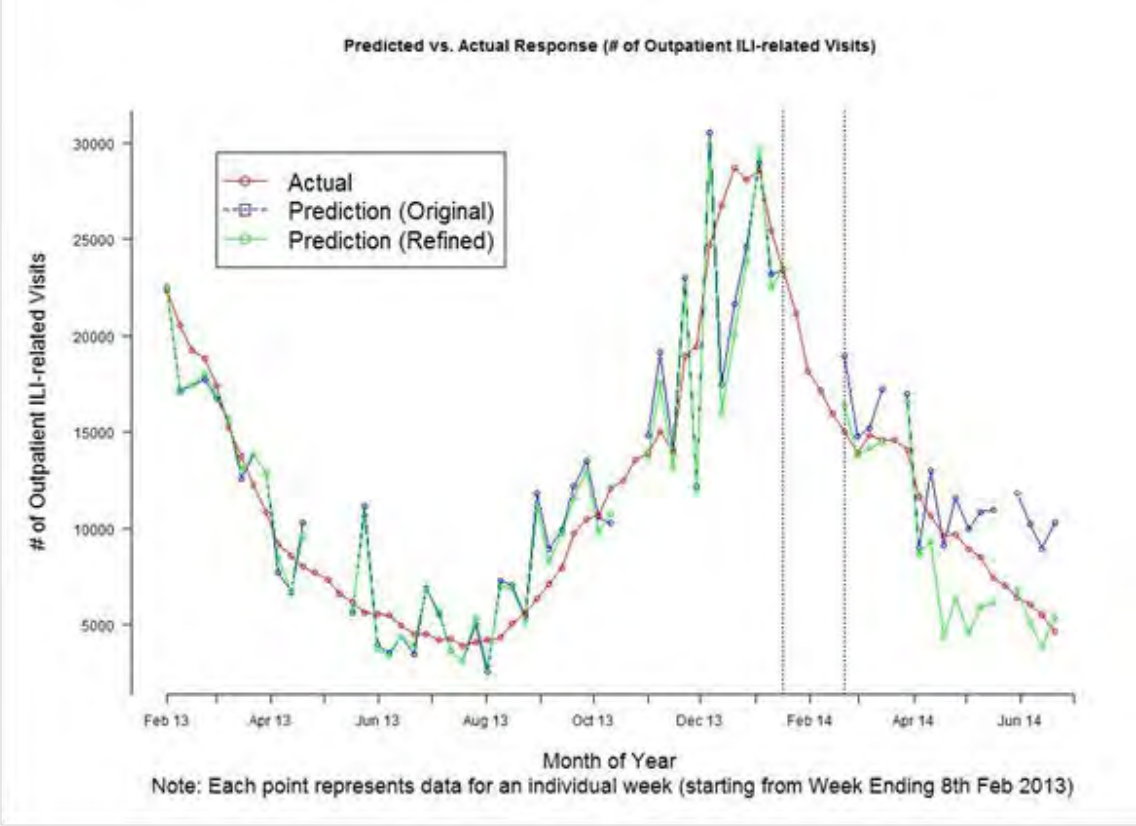

Figure 18. Predicted (Original) vs. Predicted (Refined) vs. Actual Values for Number of Outpatient ILI Visits

In summary, the high Pearson's correlation coefficient obtained for the comparison of test set predictions and actual CDC ILI Surveillance data potentially indicates a strong correlation between Twitter and CDC ILI surveillance data. In addition, the statistical summary of both models has indicated that both are well fit. For the model validation with the use of test set, predicted values of both models seem to match the

declining rate of ILI visits from February to June 2014. As the test set is small, however, it is necessary to validate the model again after obtaining a larger test set.

3. Models for HHS Regional Level

The approach did not work out well at the regional level. Table 7 shows a statistical summary of HHS regional models for predicting the number of outpatient ILI visits. Out of the ten models constructed for the ten HHS regions, only four are well fit with reasonable adj.R^2. Figure 19 shows the predicted and actual number of outpatient ILI-related visits for the best regional model (San Francisco). In general, the prediction generated from the four models does observe the up-and-down trend of the flu season despite having over-predicted by almost two-fold in a few instances. The predictions from the other six models were found to be very noisy. Figure 20 shows the predictions made for the worst regional model (Seattle). It can be observed that the prediction errors almost double or triple on most occasions.

Region	Adjusted R^2	Standard Deviation	Number of Influenza-Related Tweets
San Francisco	0.724	469.448	125084
Kansas City	0.665	161.493	42598
Chicago	0.660	455.5	259013
Atlanta	0.606	885.208	173502
New York	0.523	685.07	59812
Denver	0.497	330.255	27075
Boston	0.477	166.762	63288
Philadelphia	0.466	908.325	120773
Dallas	0.451	1401.009	134632
Seattle	0.325	115.788	49683

Table 7. Statistical Summary of HHS Regional Models that are Constructed using the Training Set for Predicting Number of Outpatient ILI Visits

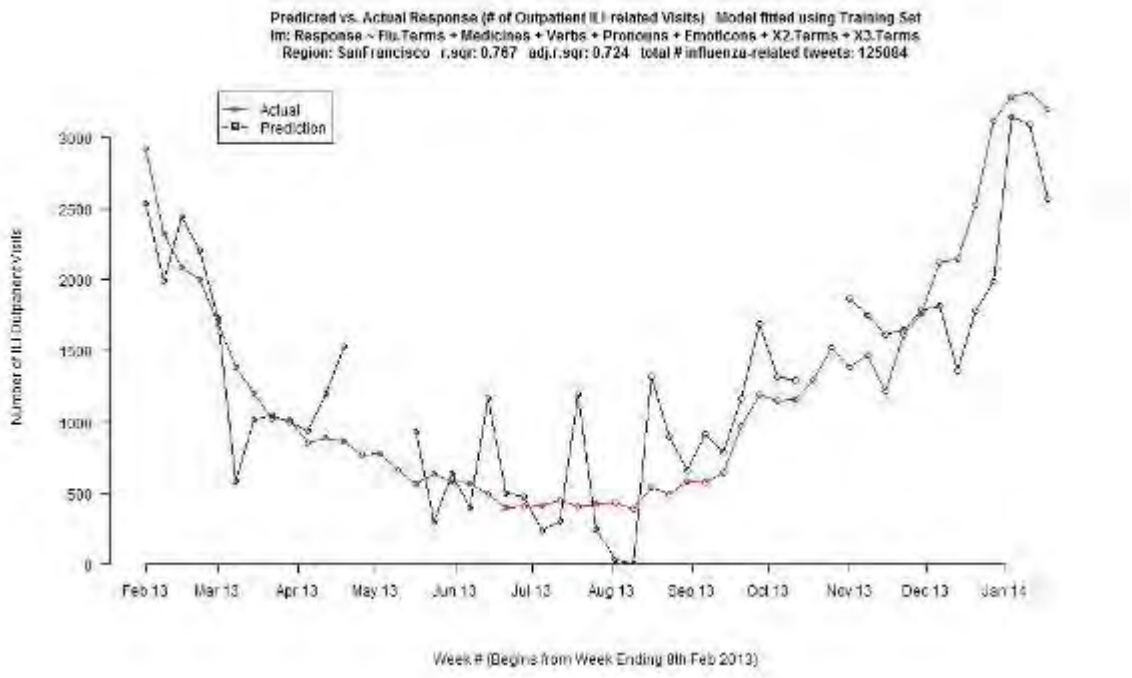

Figure 19. Predicted vs. Actual Values for Number of Outpatient ILI Visits (San Francisco)

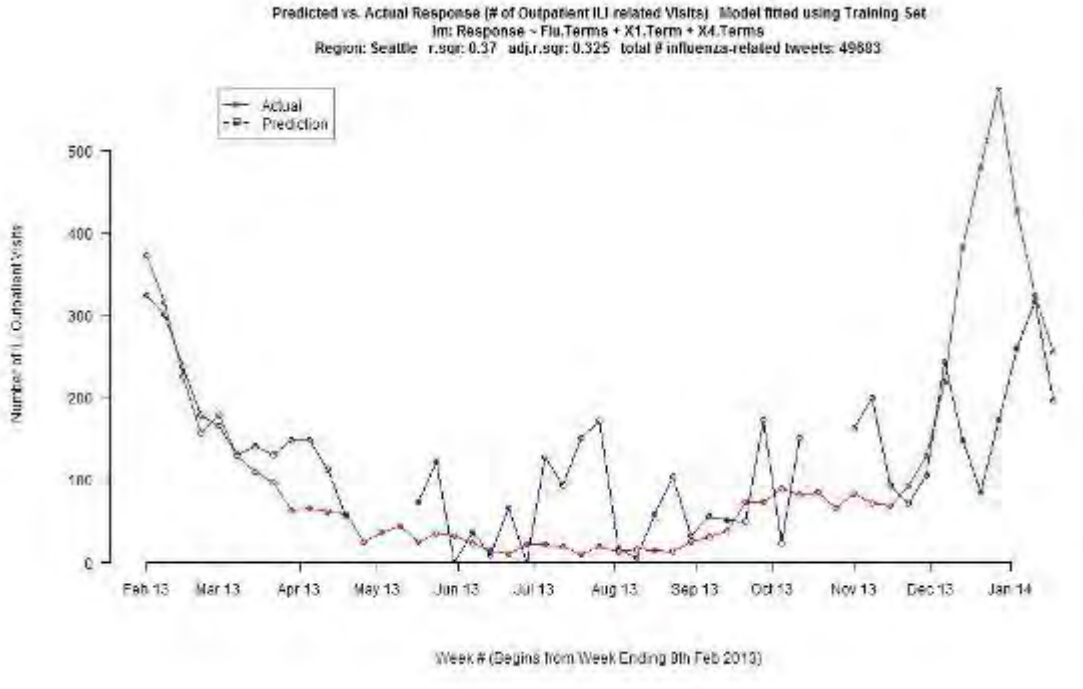

Figure 20. Predicted vs. Actual Values for Number of Outpatient ILI Visits (Seattle)

B. MODEL FOR PREDICTING NUMBER OF COLLECTED ILI RESPIRATORY SPECIMENS

The resulting model for the national level (NL) seems to suggest the presence of correlation between the Twitter messages and the number of collected ILI respiratory specimens. The NL model gives a fairly good prediction, capturing the increasing and declining trend of the number of specimens throughout the U.S. flu season.

The Pearson's correlation coefficient between the test set predictions of the NL model and actual CDC Virologic surveillance data is computed to be 0.833 (95% CI: 0.574, 0.940). After combining the training and test set predictions, the Pearson's correlation coefficient between the combined set and the actual CDC Virologic surveillance data is computed to be 0.879 (95% CI: 0.807, 0.926). Unlike the NL model, however, the models constructed for each of the 10 HHS regions have varying results. The models appears to be well fit only for one region.

1. Model for National Level

Models are constructed to predict the number of specimens collected using the best subset of predictors that is identified through the variable selection process described in Chapter IV.A.6 Fitting Models.

Figure 21 and Figure 22 show the relationship between each Indicative and Supportive predictor variable against the number of specimens. From the scatterplots, Flu.Terms, Flu.Complications and Emoticons seem to be the only predictors that are clearly correlated to the number of specimens. The trend lines does indicates that each predictor variable is positively correlated to the number of specimens with the exception of Flu.Activities, Medicines and Pronouns, which are negatively correlated, as well as Rest.Activities, which seems to form no relation with the number of specimens.

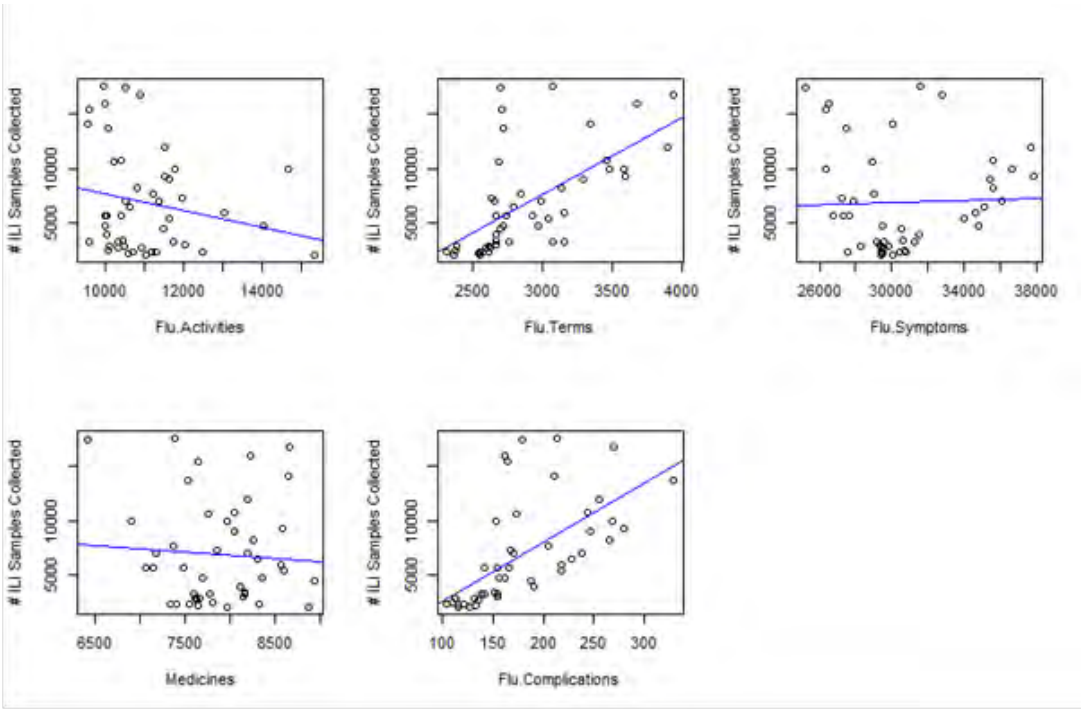

Figure 21. Relationship between Each Indicative Predictor Variable and the Number of Collected Respiratory Specimens

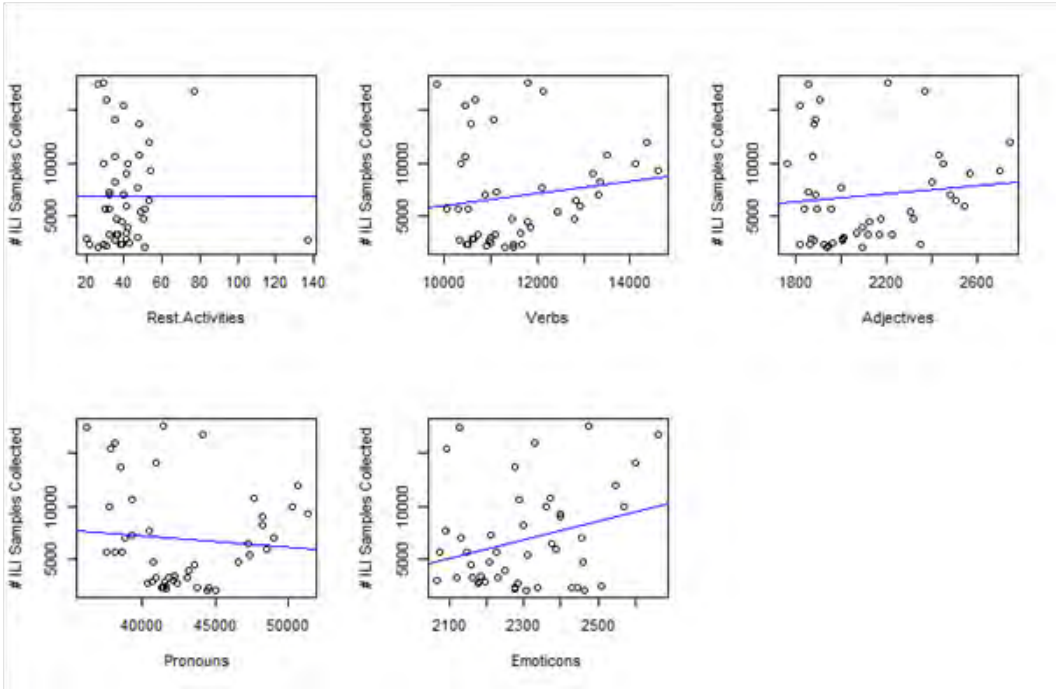

Figure 22. Relationship between Each Supportive Predictor Variable and the Number of Collected Respiratory Specimens

Table 8 shows the eight best subsets of predictors that are returned by the exhaustive search algorithm. After performing the ten-fold cross validation, the 8th subset (denoted with *) is identified as the best subset with the smallest average standard deviation of residuals of 2731.

Predictors	Subsets of Predictors							
	1	2	3	4	5	6	7	8*
Flu.Terms			*	*	*	*	*	*
Verbs						*	*	*
Adjectives							*	*
Pronouns		*						
Flu.Complications	*	*	*	*	*	*	*	*
Emoticons					*	*	*	*
Influenza.Related.Tweets						*	*	*
1.Terms				*	*			
2.Terms								*
7.Terms			*	*	*	*	*	*
Average Standard Deviation of Residuals	3779	3400	2945	2797	2886	2823	2751	2731

Table 8. Best Subsets of Predictor Variables (Original)

The model constructed using the best subset achieves a high adj.R^2 of 0.7955, which indicates that that the model is well fit. Further, a high p-Value of 4.014e-12 corresponds to a high level of confidence in terms of making predictions. Figure 23 shows the statistical summary of the constructed model, and Figure 24 the equation, for predicting the number of collected respiratory specimens.

```
Residuals:
    Min      1Q  Median      3Q     Max
-3547.9 -1164.6    65.4   853.0  6022.2

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             4968.8950  5448.6747   0.912 0.367701
Flu.Terms                  4.2199     1.1557   3.651 0.000802 ***
Verbs                      2.4844     0.9869   2.517 0.016287 *
Adjectives                 7.8688     3.6816   2.137 0.039248 *
Flu.Complications         43.7622     9.5709   4.572 5.24e-05 ***
Emoticons                 11.1086     3.3062   3.360 0.001820 **
Influenza.Related.Tweets  -0.6899     0.4831  -1.428 0.161664
X2.Terms                  -2.2761     1.7922  -1.270 0.212006
X7.Terms                 -32.9407     8.0701  -4.082 0.000229 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2141 on 37 degrees of freedom
Multiple R-squared:  0.8319, Adjusted R-squared:  0.7955
F-statistic: 22.89 on 8 and 37 DF,  p-value: 4.014e-12
```

Figure 23. Statistical Summary of Constructed Model for Number of Collected Respiratory Specimens

$$\begin{aligned}&\#\text{Collected Respiratory Specimens}\\&= 4968.9 + 4.22 \times \text{Flu.Terms} + 2.48 \times \text{Verbs} + 7.87 \times \text{Adjectives}\\&+ 43.76 \times \text{Flu.Complications} + 11.11 \times \text{Emoticons}\\&- 0.69 \times \text{Influenza.Related.Tweets} - 2.28 \times \text{X2.Terms} - 32.94 \times \text{X7.Terms}\end{aligned}$$

Figure 24. Equation for Predicting Number of Collected Respiratory Specimens

The most significant and influential predictors are Flu.Terms and Flu.Complications. This is not surprising as it has already been known that the two predictors form an increasing relationship with the number of specimens. In addition, Emoticons also appear to be statistically significant in its contribution to the number of specimens.

Another interesting observation is the negative correlation between the number of Influenza.Related.Tweets, X2.Terms and X7.Terms against response. It is already mentioned in the previous section that the number of influenza-related tweets and tweets with seven or more matching terms has decreased by two- to four-fold from the beginning of February 2013 to June 2014. This may indicate either a decrease in users'

participation in Twitter or the unintended inclusion of selected keywords that are commonly used in non-influenza-related events.

Figure 25 shows the predicted and actual number of collected respiratory specimens for the training set and test set. The prediction does capture the trend of the influenza season. As for predictions made for the test set, there are several occurrences of over-estimation but the rate of collection of specimens for the predictions did seem to decrease as per the actual values.

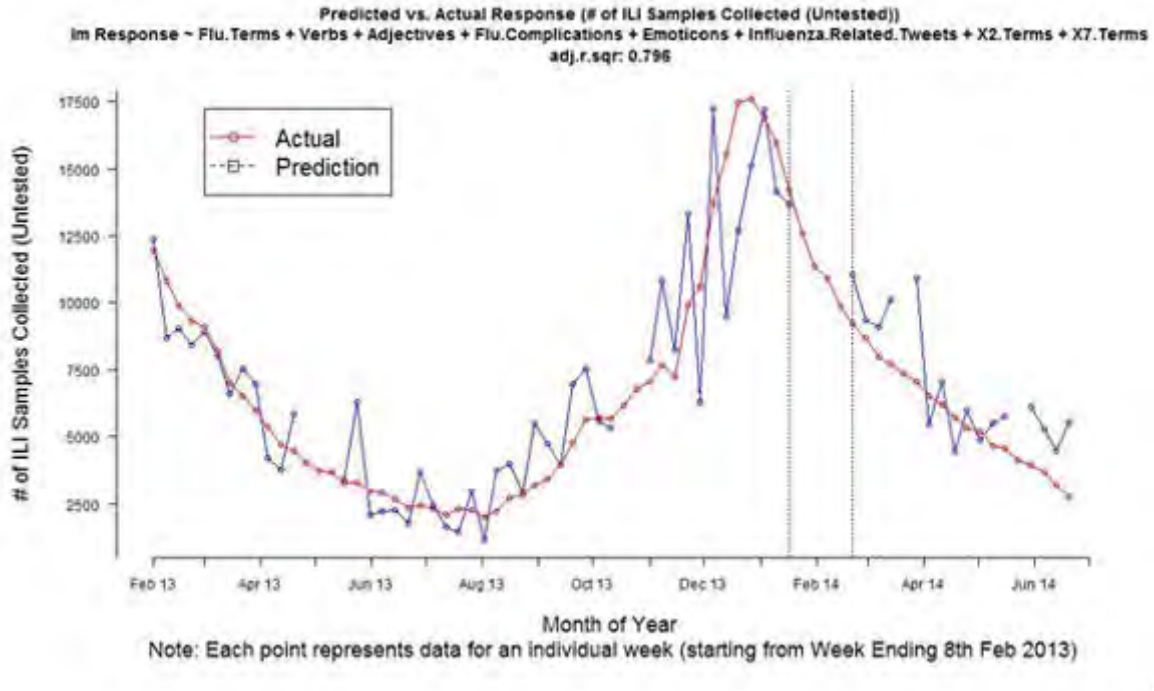

Figure 25. Predicted vs. Actual Values for Number of Collected Respiratory Specimens

While the model is able to exhibit high correlation of the predictions to the actual number of collected specimens, the standard deviation of the residuals (difference between the actual and predicted value) still seems to be too high with a 2141, which is relatively large. The plot of residuals versus predicted values in Figure 26 shows the residuals for each data point (week). There are a couple of data points with high residuals in the month of December 2013. They are further examined and found to be caused by a

sharp variation in usage of terms in Flu.Complications. In addition, potential outliers (as seen in Figure 27; data points for week ending 29/11/2013 and 13/12/2013) are observed to be caused by a high abrupt jump in the number of collected specimens in the respective weeks.

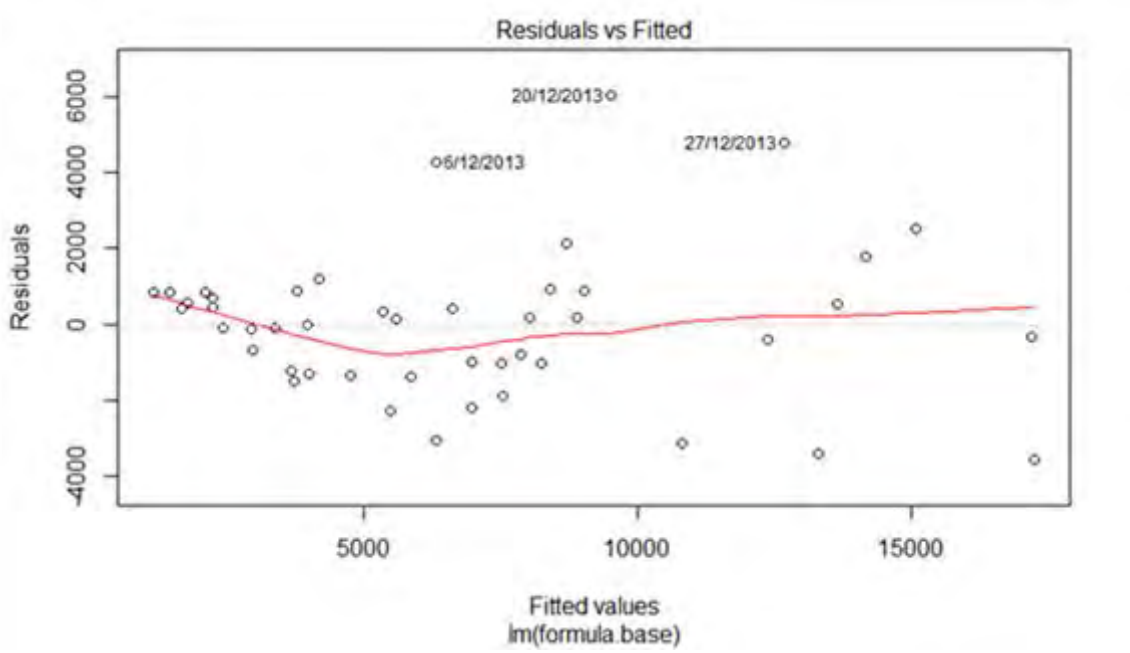

Figure 26. Residuals vs. Fitted Values for National Model

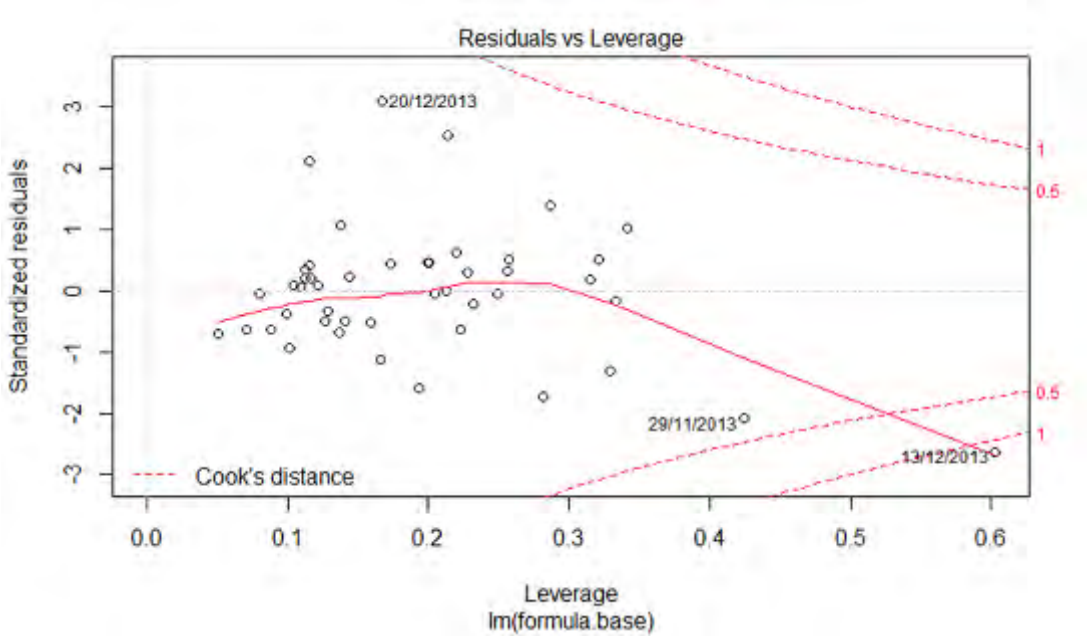

Figure 27. Residuals vs. Leverage for National Model

2. Refined Model for National Model

A refined national model is constructed with the exclusion of two outliers (data points on week ending 13/12/2013 and 20/12/2013). The model achieves an adj.R^2 of 0.852 as compared to 0.796 from the original model. In addition, its test set predictions have a higher Pearson's correlation coefficient of 0.832 (95% CI: 0.574, 0.940) as compared to the original model of 0.758 (95% CI: 0.421, 0.911). Table 9 shows the eight best subsets of predictors that are returned by the exhaustive search algorithm.

Predictors	Subsets of Predictors							
	1	2	3	4	5	6*	7	8
Flu.Terms	*				*	*	*	*
Verbs							*	*
Adjectives						*	*	*
Pronouns		*	*					
Flu.Complications		*	*	*	*	*	*	*
Emoticons			*	*	*	*	*	*
Influenza.Related.Tweets				*	*			
2.Terms						*	*	*
6.Terms								*
7.Terms				*	*	*	*	*
Average Standard Deviation of Residuals	3342	2748	2189	2170	2206	2115	2122	2129

Table 9. Best Subsets of Predictor Variables (Refined)

The same process of performing variable selection and the ten-fold cross validation returns the 6th subset (denoted with * in Table 9) as the best subset, with the smallest average standard deviation of residuals of 2115. Figure 28 shows the statistical summary for the refined model.

```
Residuals:
    Min      1Q  Median      3Q     Max
-2917.4 -1161.7  -229.6   995.8  4821.4

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       3467.9724  4573.0159   0.758 0.453042
Flu.Terms            2.6747     1.0458   2.557 0.014775 *
Adjectives           6.5994     3.0106   2.192 0.034742 *
Flu.Complications   76.7054    10.0333   7.645 3.99e-09 ***
Emoticons           13.0691     2.6456   4.940 1.70e-05 ***
X2.Terms            -3.2119     0.6393  -5.024 1.31e-05 ***
X7.Terms           -24.1215     5.7808  -4.173 0.000175 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1740 on 37 degrees of freedom
Multiple R-squared:  0.8732, Adjusted R-squared:  0.8527
F-statistic: 42.47 on 6 and 37 DF,  p-value: 3.946e-15
```

Figure 28. Statistical Summary for Refined Model

The standard deviation of the residuals of the refined model is 1740, which is smaller than 2141 from the original model. This improvement indicates the improved fit of this model. Figure 29 shows the prediction equation that is derived from the model fitted using the 6th subset of predictors.

$$\begin{aligned}\#\text{Collected.Respiratory.Specimens} = {} & 3467.97 + 2.67 \times \text{Flu.Terms} + 6.6 \times \text{Adjectives} \\ & + 76.71 \times \text{Flu.Complications} + 13.07 \times \text{Emoticons} \\ & - 3.21 \times \text{X2.Terms} - 24.12 \times \text{X7.Terms}\end{aligned}$$

Figure 29. Equation for Predicting Number of Collected Respiratory Specimens

In summary, the fit of the model improves significantly with the exclusion of the outliers. It also results in a higher Pearson's correlation coefficient obtained for the comparison between generated predictions and actual CDC Virologic Surveillance data. Figure 30 shows the actual CDC Virologic Surveillance data and the predicted values generated from the original and refined models. The refined model seems to be more precise in its test set predictions than the original model.

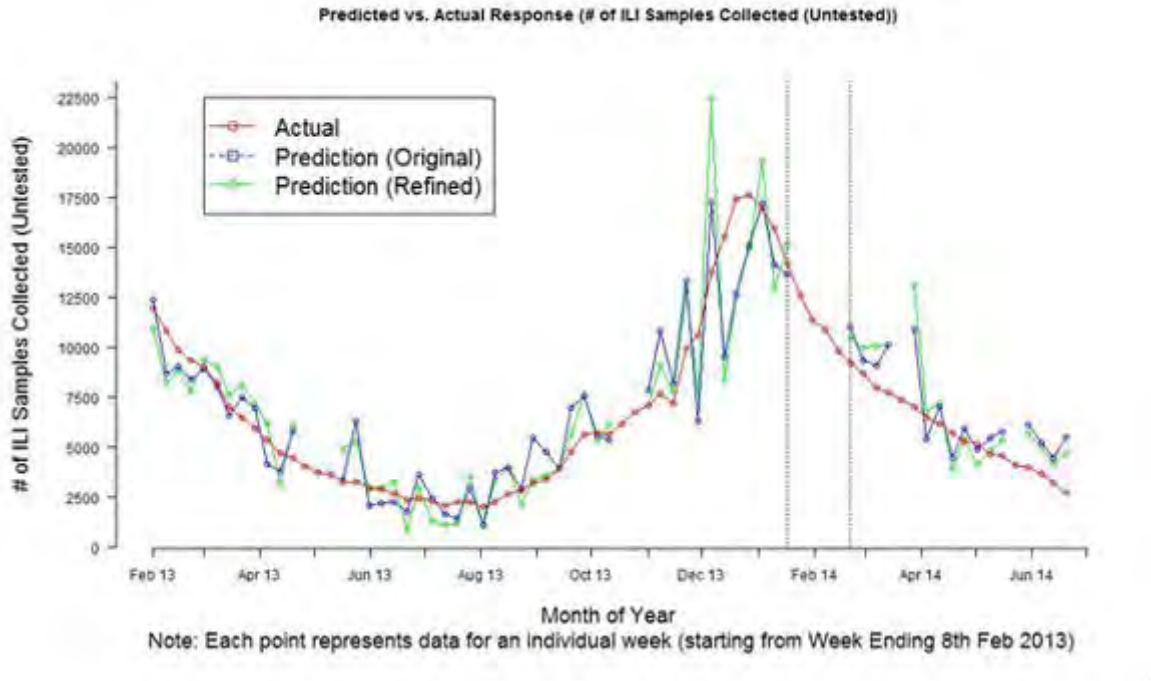

Figure 30. Predicted (Original) vs. Predicted (Refined) vs. Actual Values for Number of Collected Respiratory Specimens

3. Models for HHS Regional Level

Table 10 shows the statistical summary of HHS regional models for predicting the number of collected respiratory specimens.

Region	Adjusted R^2	Standard Deviation	Number of Influenza-Related Tweets
Kansas City	0.717	75.413	42598
San Francisco	0.635	277.849	125084
Atlanta	0.615	520.613	173502
Chicago	0.585	238.451	259013
New York	0.507	333.003	59812
Seattle	0.455	175.887	49683
Philadelphia	0.413	337.769	120773
Denver	0.413	589.517	27075
Dallas	0.378	800.268	134632
Boston	0.365	144.762	63288

Table 10. Statistical Summary of HHS Regional Models that are Constructed using the Training Set for Predicting Number of Collected Respiratory Specimens

The approach did not work out well at the regional level. Out of the ten models constructed for the ten HHS regions, only three are well fit with reasonable coefficient of determination (R^2). Figure 31 shows the predicted and actual number of collected respiratory specimens for the best regional model (Kansas City). In general, the prediction generated from the three models does observe the up-and-down trend of the flu season despite having over-predicted (under-predicted) by almost two-fold in a few instances. The predictions from the other seven models were found to be inaccurate on most occasions. Figure 32 shows the predictions made for the worst regional model (Boston).

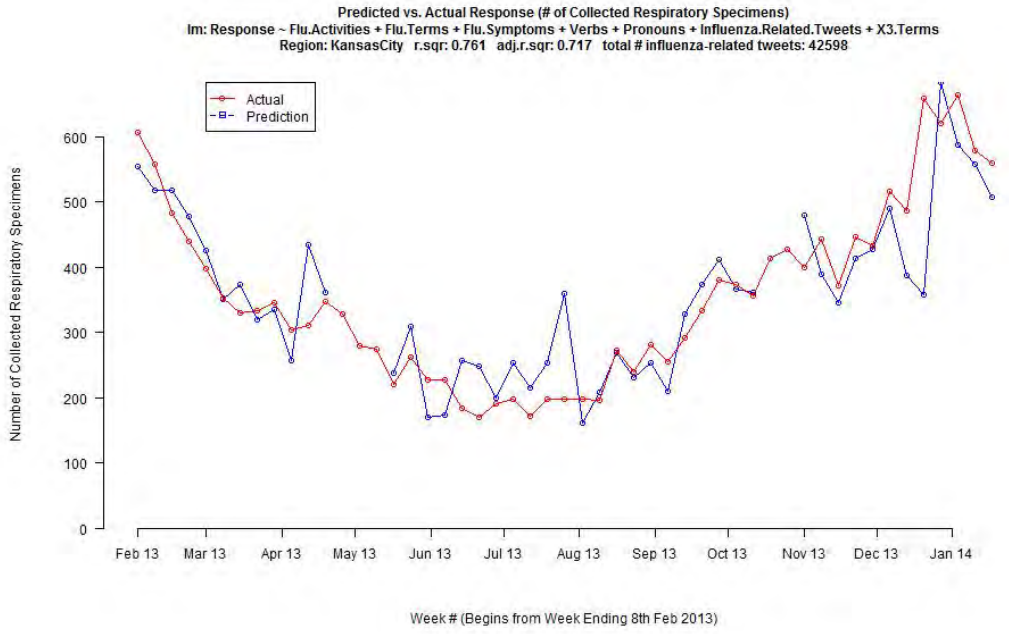

Figure 31. Predicted vs. Actual Values for Number of Collected Respiratory Specimens for Kansas City

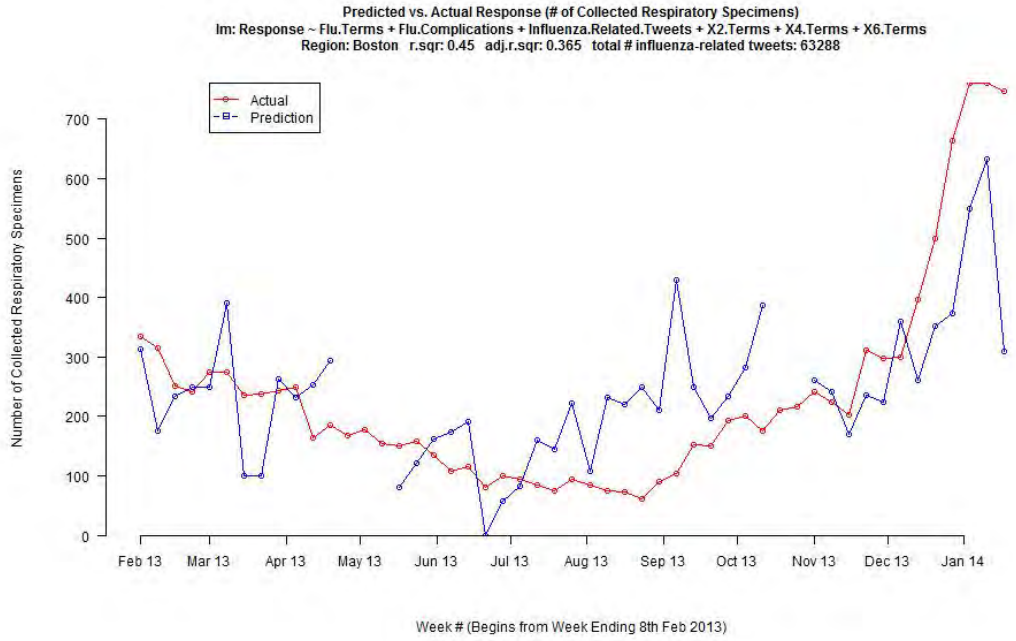

Figure 32. Predicted vs. Actual Values for Number of Collected Respiratory Specimens for Boston

C. MODEL FOR PREDICTING NUMBER OF RESPIRATORY SPECIMENS TESTED POSITIVE FOR INFLUENZA TYPE A OR B

This section discusses the results obtained from the models constructed to predict the number of respiratory specimens that are tested positive for influenza type A or B (positive specimens). At the national level, the best NL model did not perform as well as the models constructed for predicting the number of outpatient ILI visits and the number of collected respiratory specimens. It does capture the increasing and declining trend throughout the U.S. flu season but fails in its precision of prediction. The Pearson's correlation coefficient between the overall (combined training and test set) predictions of the refined NL model and actual CDC Virologic surveillance data is computed to be 0.81 (95% CI: 0.70, 0.88). While the overall predictions appears satisfactory in terms of their correlation with CDC data, the Pearson's correlation coefficient between the test set predictions and actual CDC Virologic surveillance data is computed to be only 0.613 (95% CI: 0.168, 0.850). At the regional level, the models appear to be well fit only for one out of ten regions.

1. Model for National Level

Figure 33 and Figure 34 show the relationship between each Indicative and Supportive predictor variable against the number of positive specimens. From the scatterplots, Flu.Terms, Flu.Complications and Emoticons seem to be the only predictors that are positively correlated to the number of positive specimens. The rest of the predictor variables are either negatively correlated or have no relationship with the number of positive specimens.

Table 11 shows the eight best subsets of predictors that are returned by the exhaustive search algorithm. After performing the ten-fold cross validation, the 6th subset (denoted with *) is identified as the best subset with the smallest average standard deviation of residuals of 1009.

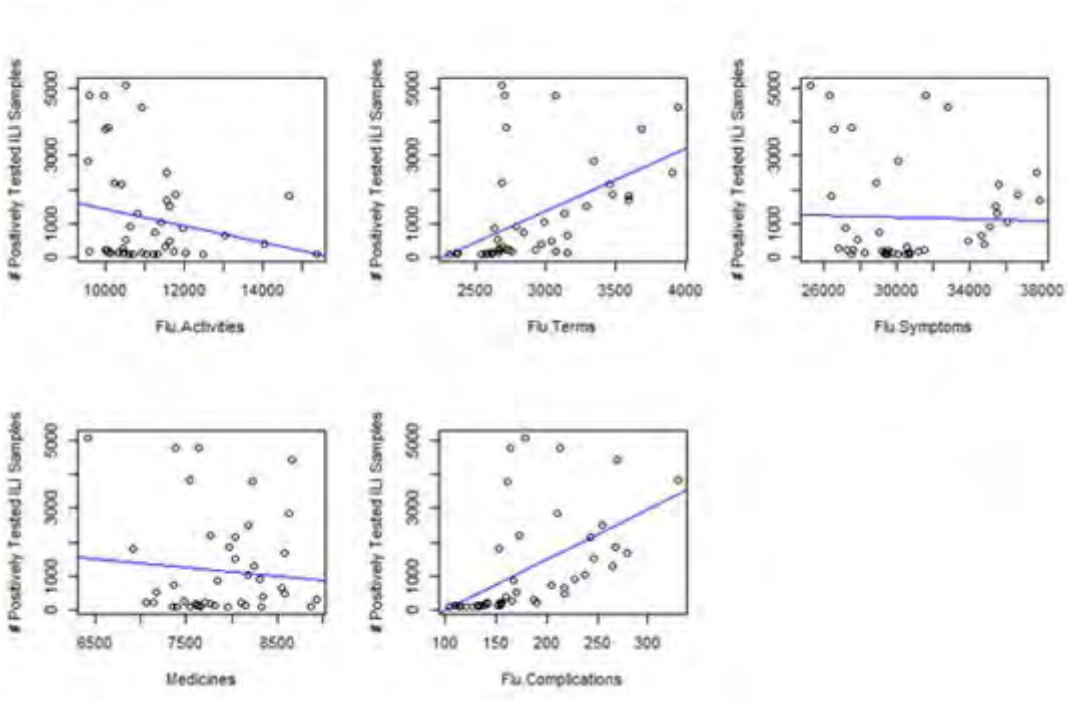

Figure 33. Relationship between Each Indicative Predictor Variable and the Number of Respiratory Specimens Tested Positive

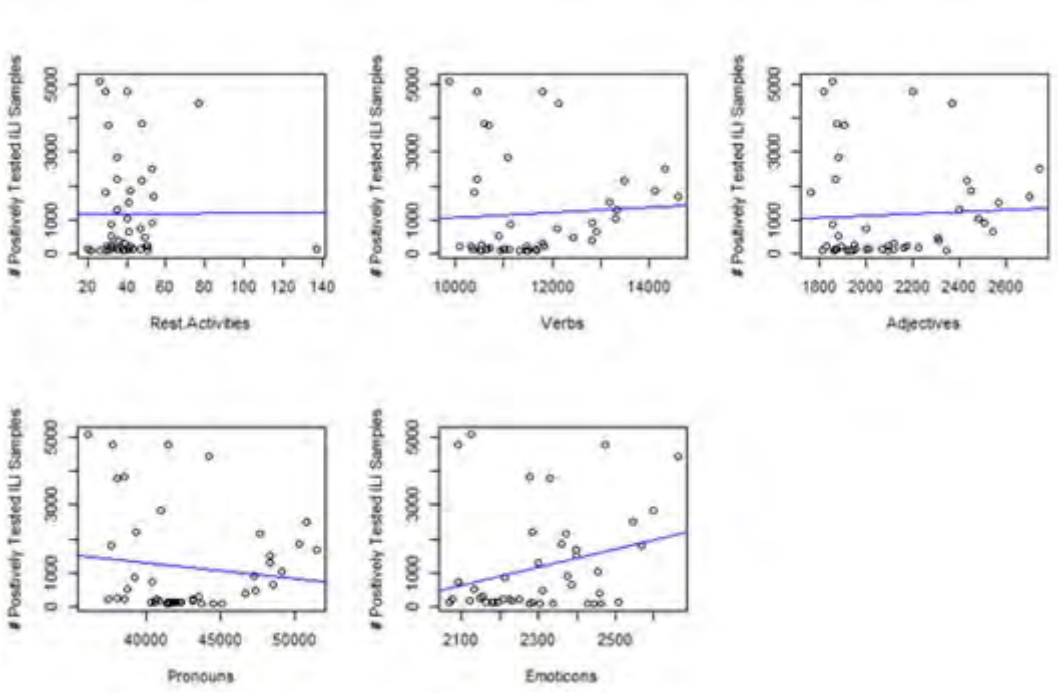

Figure 34. Relationship between Each Supportive Predictor Variable and the Number of Respiratory Specimens Tested Positive

Predictors	Subsets of Predictors							
	1	2	3	4	5	6*	7	8
Flu.Activities								*
Flu.Terms			*	*	*	*	*	*
Adjectives						*	*	*
Pronouns							*	
Flu.Complications	*	*	*	*	*	*	*	*
Emoticons					*	*	*	*
Influenza.Related.Tweets							*	*
1.Terms				*	*	*		
5.Terms								*
7.Terms		*	*	*	*	*	*	*
Average Standard Deviation of Residuals	1296	1134	1049	1020	1028	1009	1013	1016

Table 11. Best Subsets of Predictor Variables (Original)

The model constructed using the best subset achieves a fairly high adj.R^2 of 0.6799, which indicates that that the model is fairly well fit. Figure 35 shows the statistical summary of the constructed model, and Figure 36 the equation, for predicting the number of respiratory specimens tested positive.

```
Residuals:
     Min       1Q   Median       3Q      Max
-1601.97  -388.21   -42.36   407.18  2640.96

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       2013.0586  2158.5310   0.933 0.356762
Flu.Terms            1.1886     0.4487   2.649 0.011594 *
Adjectives           2.1836     1.2332   1.771 0.084430 .
Flu.Complications   16.0690     3.3665   4.773 2.56e-05 ***
Emoticons            2.7748     1.0741   2.583 0.013649 *
X1.Term             -0.5591     0.1637  -3.416 0.001499 **
X7.Terms           -11.1488     2.6950  -4.137 0.000182 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 851.6 on 39 degrees of freedom
Multiple R-squared:  0.7226, Adjusted R-squared:  0.6799
F-statistic: 16.93 on 6 and 39 DF,  p-value: 1.649e-09
```

Figure 35. Statistical Summary of Constructed Model for Number of Respiratory Specimens Tested Positive

$$\begin{aligned}\#\text{Respiratory.Specimens.Tested.Positive} = {} & 2013.06 + 1.19 \times \text{Flu.Terms} + 2.18 \times \text{Adjectives} \\ & + 16.07 \times \text{Flu.Complications} + 2.77 \times \text{Emoticons} \\ & - 0.56 \times \text{X1.Term} - 11.15 \times \text{X7.Terms}\end{aligned}$$

Figure 36. Equation for Predicting Number of Respiratory Specimens Tested Positive

The most significant and influential predictors are Flu.Complications and X7.Terms. This is not surprising as it has already been known that the Flu.Complications forms an increasing relationship with the number of positive specimens. Another interesting observation is the negative correlation between the number of X1.Terms and X7.Terms against response. It is already mentioned in the previous section that the tweets with seven or more matching terms have massively decreased by two- to four-fold from the beginning of February 2013 to June 2014. This massive decrease is also observed for

X1.Term. This may indicate either a decrease in users' participation in Twitter or the unintended inclusion of selected keywords that are commonly used in non-influenza-related events.

Figure 37 shows the predicted and actual number of positive specimens for the training set and test set. The predictions are mostly inaccurate even though the predicted values did rise to match the high number of positive specimens during the peak of the influenza season.

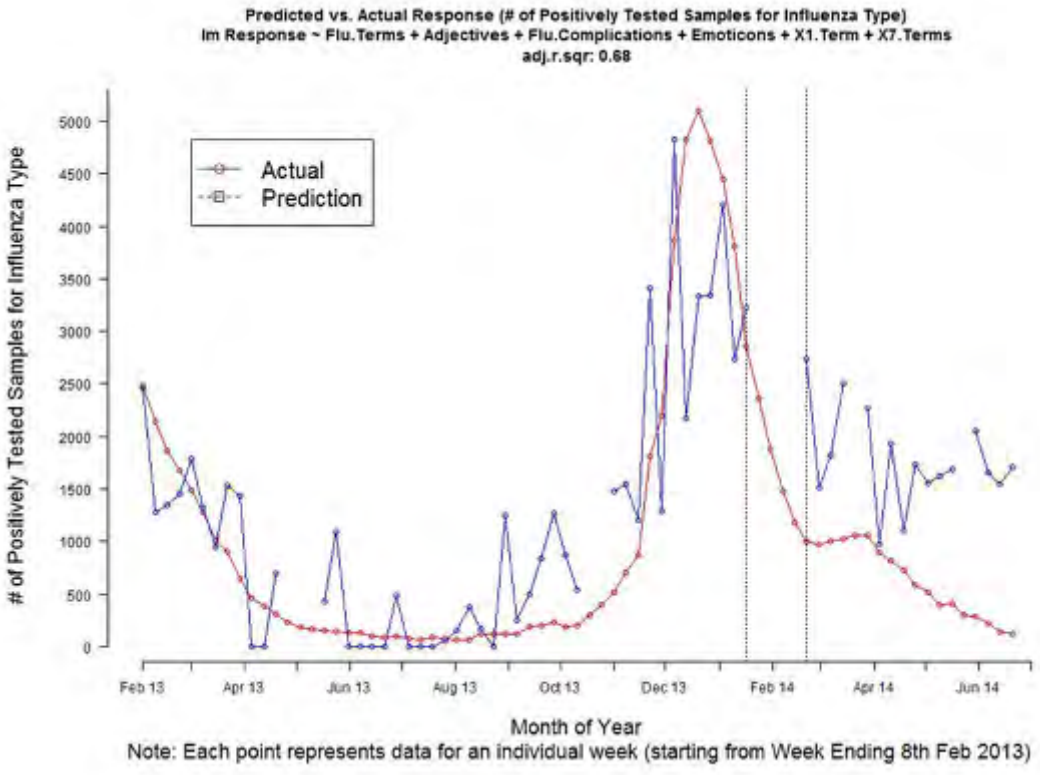

Figure 37. Predicted vs. Actual Number of Respiratory Specimens Tested Positive for Influenza Type A or B

2. Refined Model for National Model

The refined model achieves a R^2 of 0.77 as compared to 0.68 of the original model. In addition, its overall predictions (combined training and test set) has a higher

Pearson's correlation coefficient of 0.804 (95% CI: 0.693, 0.877) as compared to the original model of 0.766 (95% CI: 0.639, 0.852).

The standard deviation of the residuals is 661.6, which is much smaller than 851.6 from the original model. This improvement also indicates the improved fit of this model. Figure 38 shows the statistical summary of the refined model and Figure 39 the prediction equation that is derived from the model constructed using the 6th subset of predictors.

Figure 40 compares the actual CDC Virologic Surveillance data and the predicted values generated from the original and refined models. The refined model seems to be more precise in its prediction than the original model.

```
Residuals:
    Min      1Q  Median      3Q     Max
-1368.0  -346.1  -102.0   299.4  2066.3

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       -872.0385  1837.5370  -0.475  0.63788
Flu.Terms            1.3230     0.4167   3.175  0.00302 **
Flu.Complications   19.8890     3.6964   5.381 4.33e-06 ***
Emoticons            3.8217     0.8640   4.423 8.24e-05 ***
X1.Term             -0.3561     0.1033  -3.448  0.00143 **
X7.Terms           -10.7023     1.9320  -5.539 2.64e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 661.6 on 37 degrees of freedom
Multiple R-squared:  0.7967, Adjusted R-squared:  0.7692
F-statistic:    29 on 5 and 37 DF,  p-value: 7.6e-12
```

Figure 38. Statistical Summary of Refined Model for Number of Respiratory Specimens Tested Positive

$$\begin{aligned}\#\text{Respiratory.Specimens.Test.Positive} = {} & -872.04 + 1.32 \times \text{Flu.Terms} \\ & + 19.89 \times \text{Flu.Complications} \\ & + 3.82 \times \text{Emoticons} - 0.36 \times \text{X1.Term} \\ & - 10.7 \times \text{X7.Terms}\end{aligned}$$

Figure 39. Equation for Predicting Number of Respiratory Specimens Tested Positive

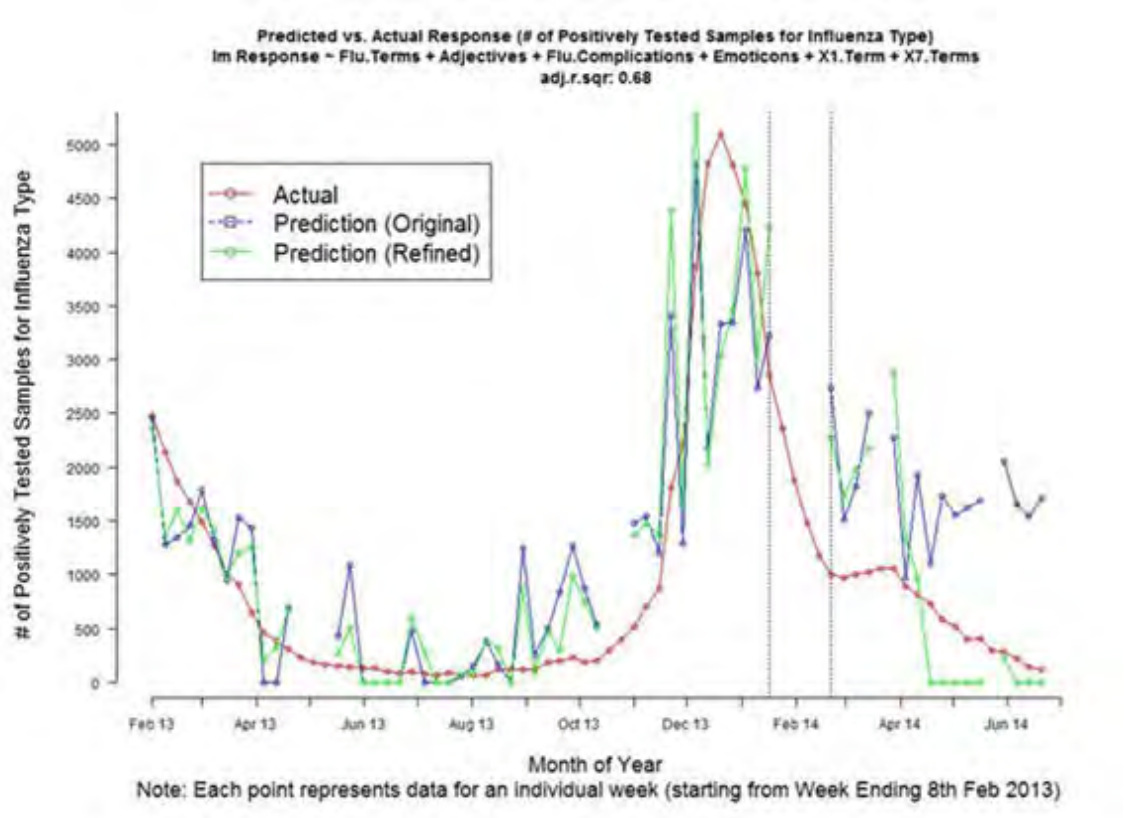

Figure 40. Predicted (Original) vs. Predicted (Refined) vs. Actual Values for Number of Respiratory Specimens Tested Positive for Influenza Type A or B

3. Models for HHS Regional Level

Table 12 shows the statistical summary of HHS regional models for predicting the number of respiratory specimens tested positive for influenza type A or B.

Region	Adjusted R^2	Standard Deviation	Number of Influenza-Related Tweets
Chicago	0.588	92.795	259013
Kansas City	0.567	35.143	42598
San Francisco	0.538	127.241	125084
Atlanta	0.454	154.572	173502
Philadelphia	0.394	134.172	120773
Seattle	0.370	83.856	49683
Denver	0.354	196.07	27075
New York	0.318	94.702	59812
Boston	0.301	63.105	63288
Dallas	0.245	279.191	134632

Table 12. Statistical Summary of HHS Regional Models that are Constructed using the Training Set for Predicting Number of Respiratory Specimens Tested Positive for Influenza Type A or B

The approach did not work out well at the regional level. None of the ten regional models are constructed to a reasonable adj.R^2. Figure 41 shows the predicted and actual number of positive specimens for the best regional model (Chicago). The Chicago predictions are found to be inaccurate on most occasions, overestimating when there are no cases of positive specimens.

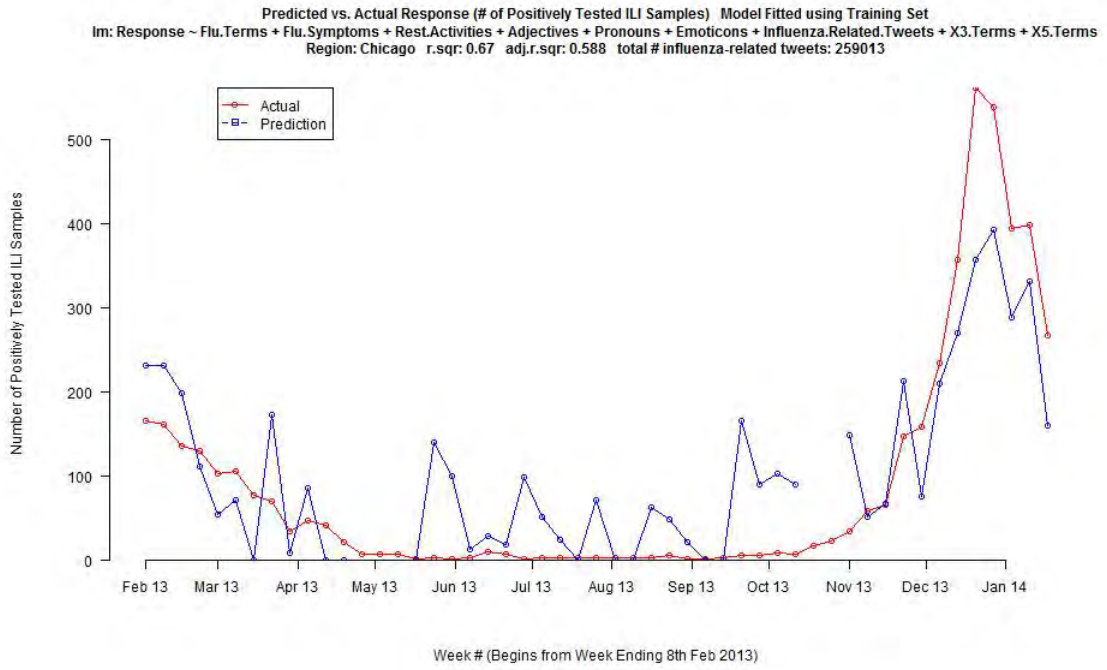

Figure 41. Predicted vs. Actual Values for Number of Respiratory Specimens Tested Positive for Influenza Type A or B for Chicago

D. MODEL FOR PREDICTING NUMBER OF INFLUENZA-ASSOCIATED HOSPITALIZATIONS

This section discusses the models that are constructed to predict the number of influenza-associated hospitalizations for the ten states selected for influenza-associated hospitalization surveillance. Accurate predictions of a hike in influenza-associated hospitalizations would help to increase response time and improve preparations to face a potential flu pandemic.

However, the constructed models did not provide predictions that are good enough for practical use. Table 13 shows the statistical summary of each constructed regional models. Out of the 12 models, only four are well fit with reasonable coefficient of determination (R^2). Figure 42 shows the predictions made by the model (Maryland) with the best fit among the ten states. The Maryland model achieves a high adj.R^2 of 0.775 but its predictions are way off of the actual values. Refer to Appendix B.4 to view the predicted versus actual values plots for the other nine states.

Region	Adjusted R^2	Standard Deviation	Number of Influenza-Related Tweets
Maryland	0.775	10.038	12014
Tennessee	0.652	6.98	11717
Colorado	0.636	9.381	9163
Georgia	0.625	7.618	26281
Oregon	0.575	9.2	9534
California	0.535	10.314	53401
New Mexico	0.493	9.129	1724
Minnesota	0.455	16.932	9744
Connecticut	0.408	30.59	7896
Michigan	0.419	8.955	21304
Ohio	0.398	11.95	25137
Utah	0.275	15.625	3376

Table 13. Statistical Summary of Models that are Constructed using the Training Set for Predicting the Rate of Influenza-Associated Hospitalizations

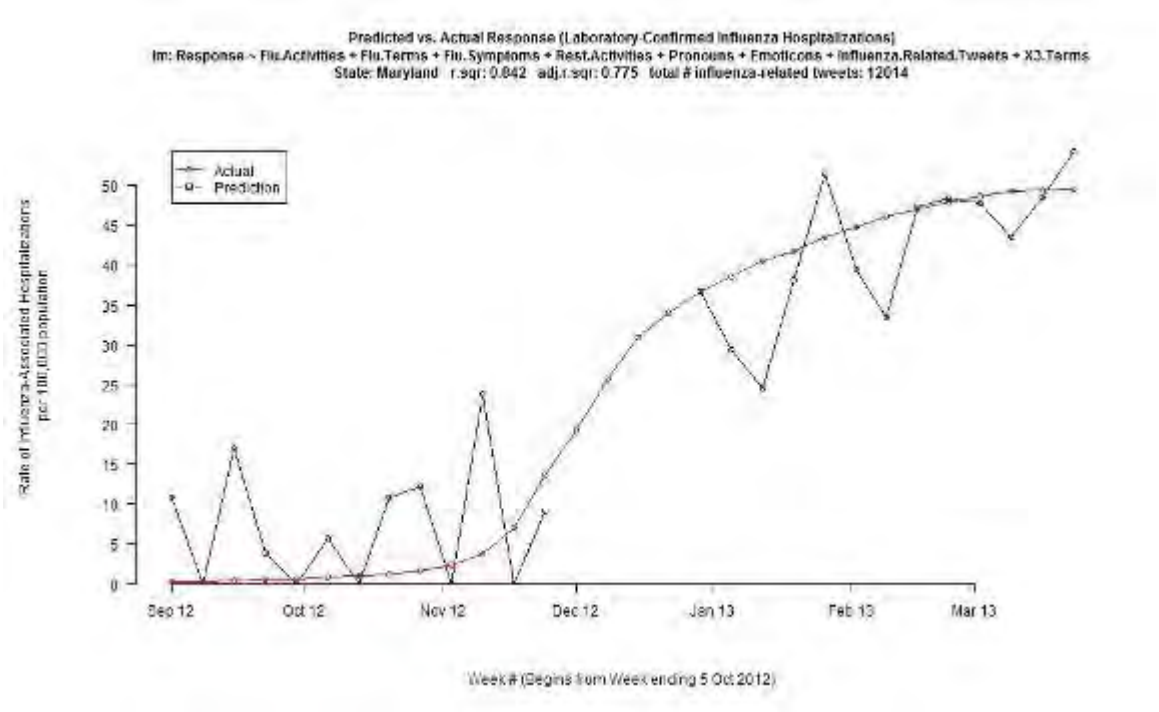

Figure 42. Predicted vs. Actual Values for Number of Influenza-Associated Hospitalizations for Maryland

Based on the resulting fit of the models, it can be concluded that the approach did not work or is infeasible for predicting the number of influenza-associated hospitalizations by state. The poor fit could be a result of keywords that are selected with the intention of capturing influenza activities instead of hospitalizations. Hence, a potential future work would be to refine the keyword selection process to acquire a dataset that is more suitable for predicting influenza-associated hospitalizations.

VI. CONCLUSIONS

A. SUMMARY

The study has strengthened the claim that Twitter is a potential indicator of influenza activity level. The exploration of using counts obtained for various categories of keywords, such as flu symptoms, verbs and medicines against the actual CDC surveillance data seems to be successful. At the national level, the constructed models are able to provide a good weekly estimate of influenza activity indicators such as number of ILI outpatient visits and number of collected respiratory specimens.

The models' predictions match the increasing and declining trend across the U.S. influenza season. The Pearson's correlation coefficient between the test set predictions of the number of outpatient ILI visits using the constructed model and actual CDC ILI surveillance data is computed to be 0.900 (95% CI: 0.732, 0.965). The same approach, however, only succeeds for a subset of regional models.

There are a few ways to further improve the models to achieve better results. Firstly, the adopted simple and random keyword selection process could certainly be changed to select keywords that are more relevant to filter out non-influenza-related tweets. Secondly, the research is conducted using just 1% of the entire tweets each day, and this may have an impact on the fit of the models.

In the long run, the change in social behavior and lifestyle will result in a need for the adaptation of the current models. This is evidently proven as even Google Flu Trends (GFT) sees a need to update their initial 2009 model in October 2013 (Google 2014).

Finally, we ask whether Twitter can predict the influenza activity level in the U.S. The answer appears to be yes. The national-level models have proven to be highly correlated with CDC's ILI and Virologic Surveillance data. This outcome of this research has certainly helped in gaining optimism to pursue a more extensive study.

B. RECOMMENDED FUTURE WORK

This section lists the future research that could be implemented to improve the proposed approach.

(1) Geo-location Prediction

The omission of influenza-related tweets due to the anonymity of the user's location could have an effect on the resulting models constructed for the HHS regions. In this study, the locale of a tweet originator is determined by comparing the text of the tweet's location field against the names of each U.S. state as well as 50 major cities. This approach manages to identify the state for 47% of the 375,000 influenza-related tweets.

Twitter users have the option to tag each tweet with their current geo-location (Stone 2009). However, this geo-location option is seldom activated (Evans 2010). This lack of location information has led to the research and development of geo-location prediction tools based on the user's past tweets or social networks. An accurate geo-location prediction tool will help to determine the locale information, such as the particular state or city in which the Twitter user is residing. Furthermore, this will potentially increase the number of influenza-related tweets for the regional models.

(2) Refined Keyword Selection

The current method of keyword selection is both simple and manual. This has potentially led to the incorrect classification of influenza-related tweets. An alternative manual method would be to deploy a human annotator to identify influenza-related tweets followed by picking up the most frequently used keywords in these tweets.

This alternative method could be further improved by giving different weightage to keywords. The adjective 'sick' is commonly used to declare one's ailing or ill health. It could also be used to declare one's negative feeling about certain issues happening. Hence, by assigning a lower weightage to matching keywords that raise uncertainty of an influenza-related event, the impact of an incorrect classification could be reduced.

(3) Application of Approach for Predicting Level of Influenza Activity in Other Countries

The World Health Organization (WHO) keeps track of various influenza activity statistics, such as the number of influenza viruses detected, with the influenza surveillance network, FluNet (WHO 2014). The statistics are consolidated with the collaboration of worldwide National Influenza Centres who collect and provide virus specimens to WHO Collaborating Centres. The presence of correlation between U.S. tweets and CDC ILI & Virologic Surveillance data certainly indicate the possibility of correlation between the tweets and the influenza activity statistics of another country.

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX A. PLOTS

1. Predicted vs. Actual Number of Outpatient ILI Visits (Regional)

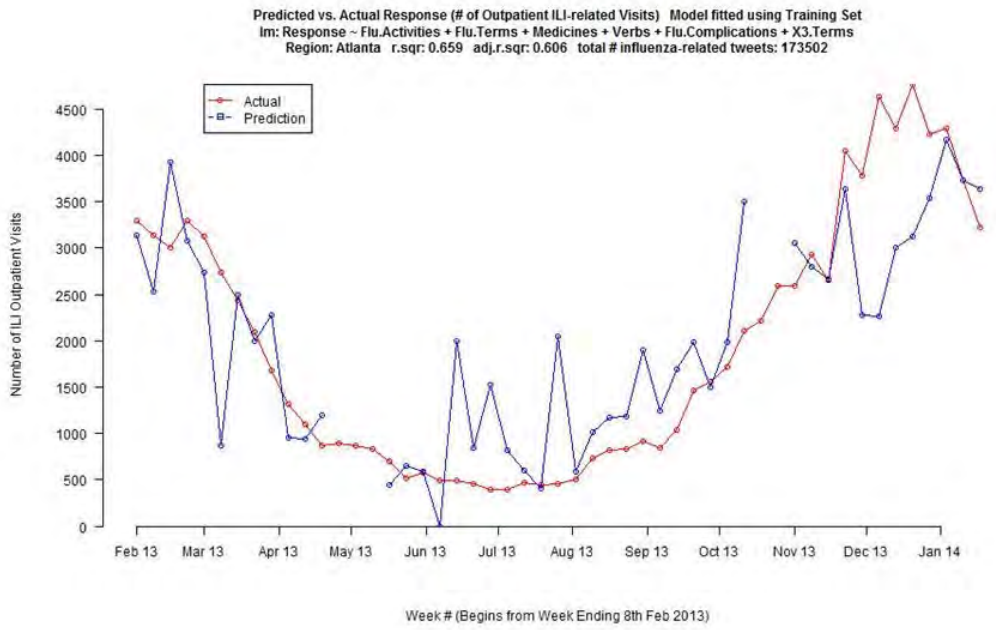

Figure 43. Predicted vs. Actual Number of Outpatient ILI Visits (Atlanta)

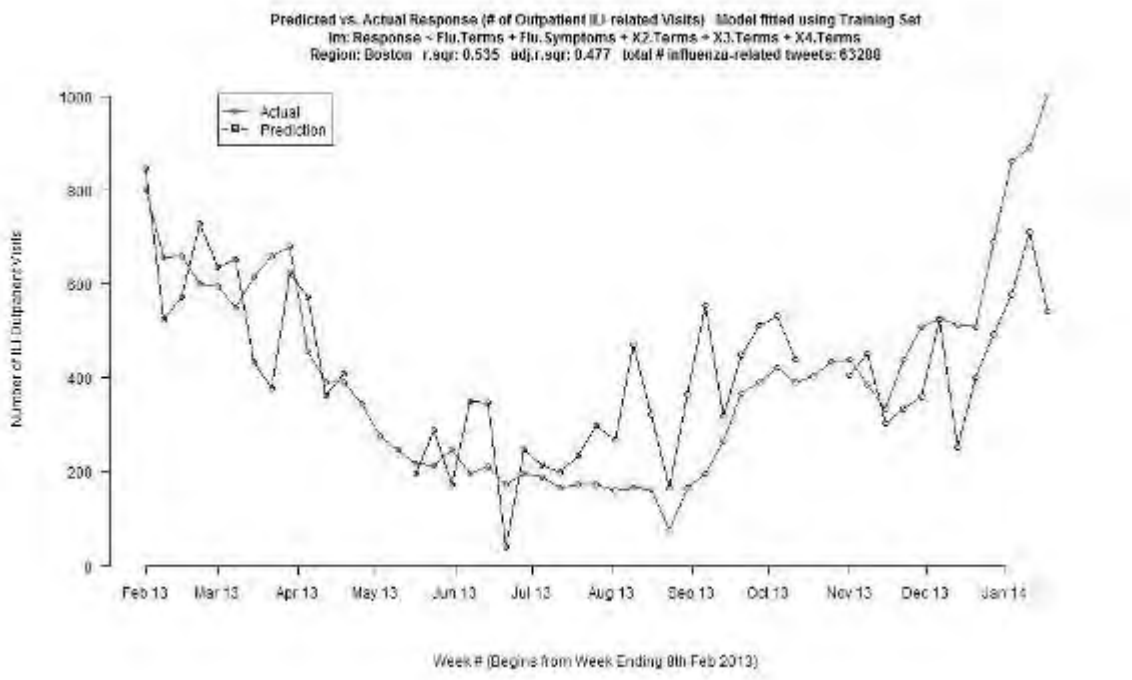

Figure 44. Predicted vs. Actual Number of Outpatient ILI Visits (Boston)

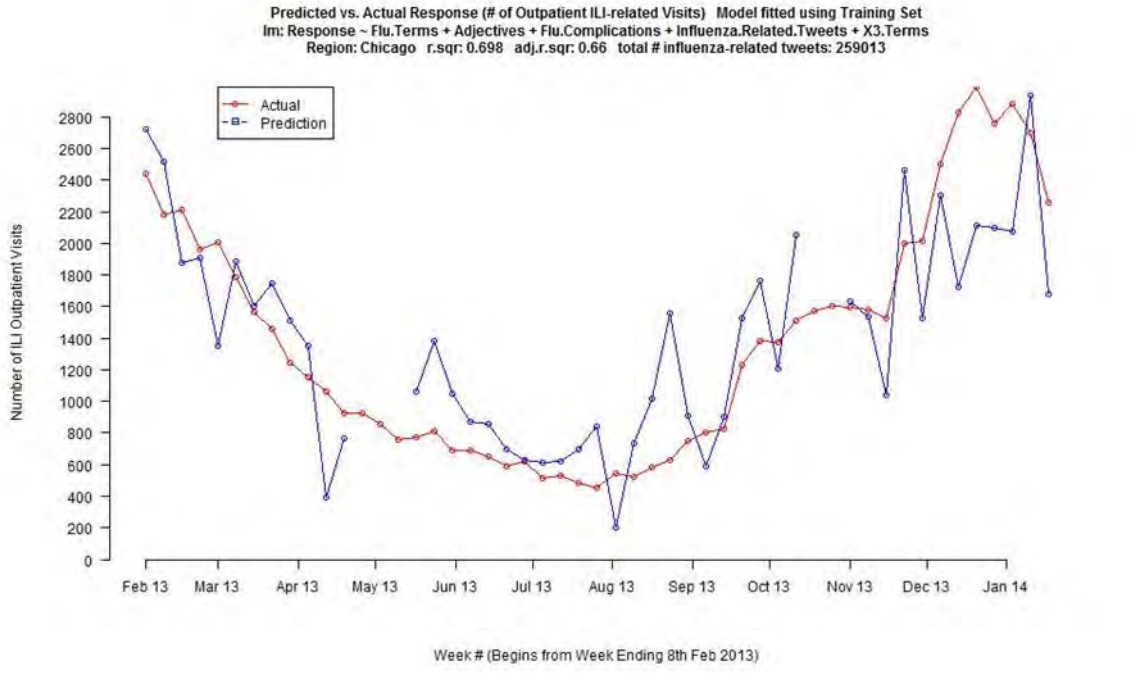

Figure 45. Predicted vs. Actual Number of Outpatient ILI Visits (Chicago)

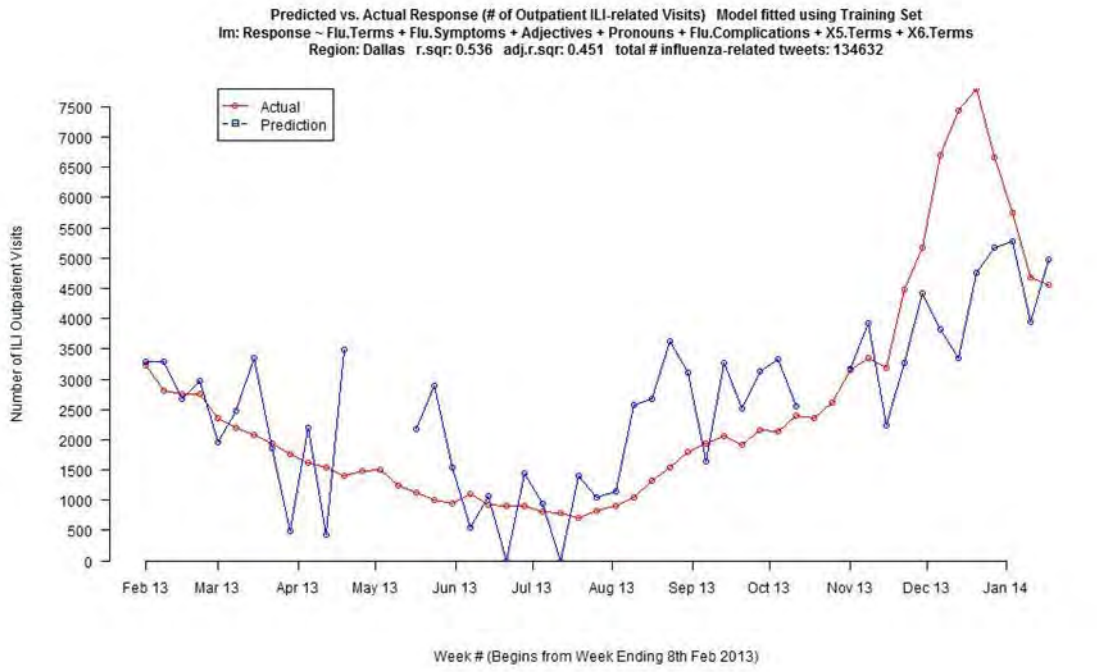

Figure 46. Predicted vs. Actual Number of Outpatient ILI Visits (Dallas)

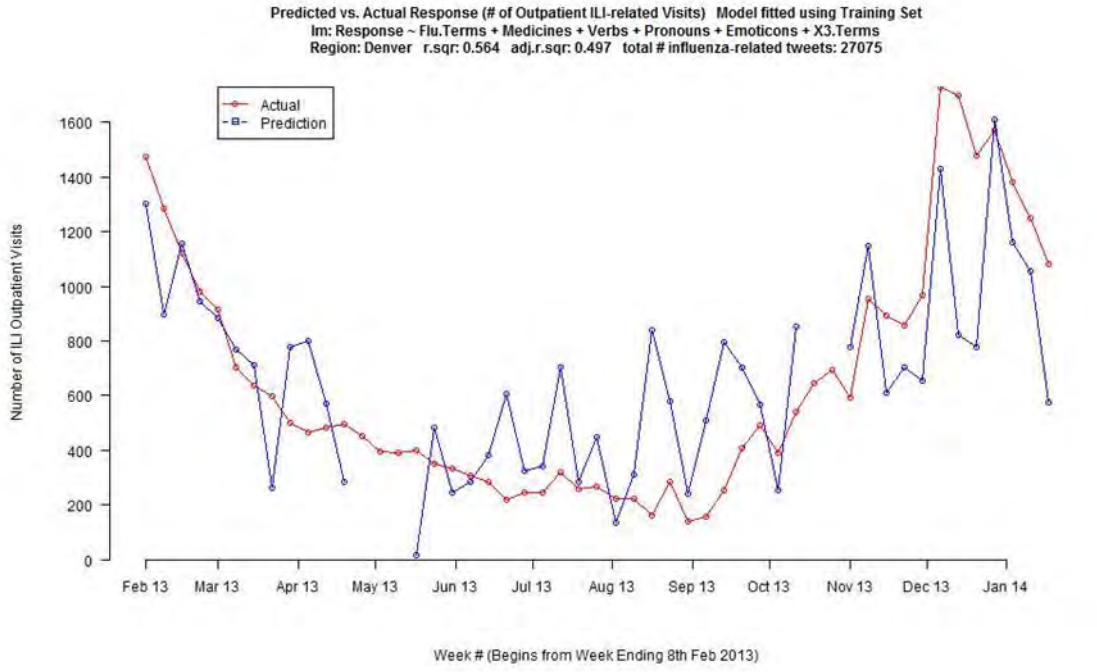

Figure 47. Predicted vs. Actual Number of Outpatient ILI Visits (Denver)

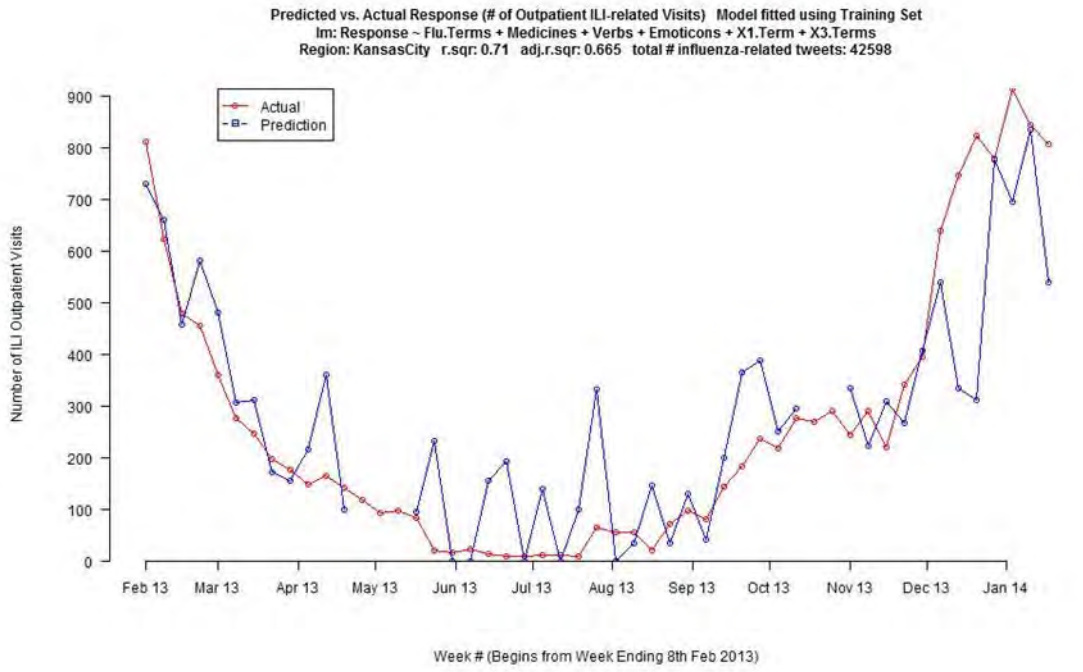

Figure 48. Predicted vs. Actual Number of Outpatient ILI Visits (Kansas City)

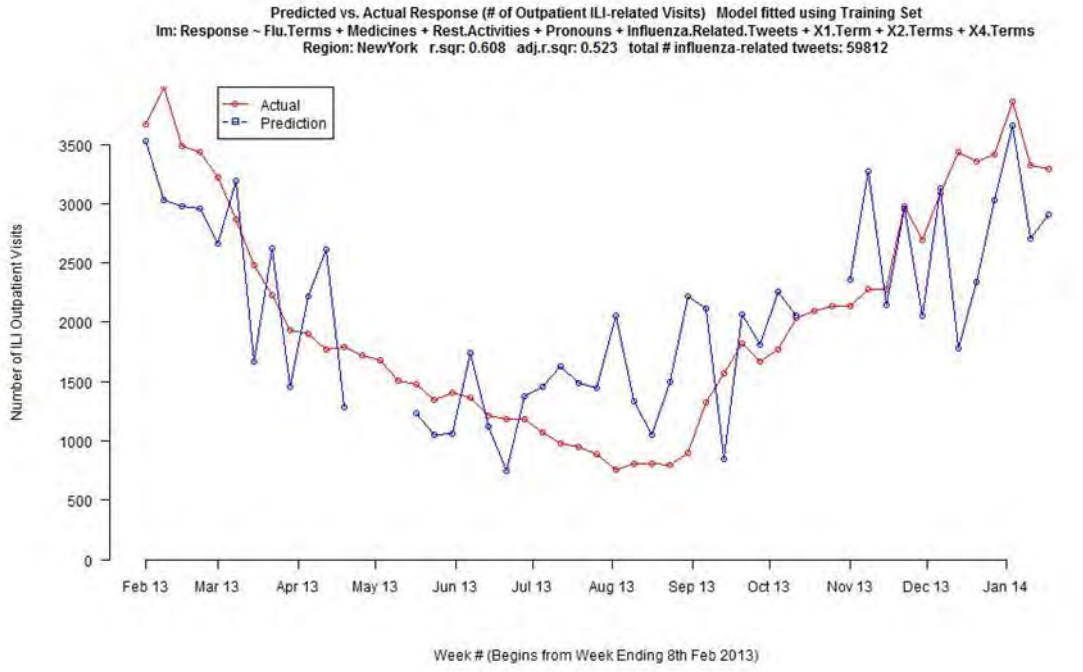

Figure 49. Predicted vs. Actual Number of Outpatient ILI Visits (New York)

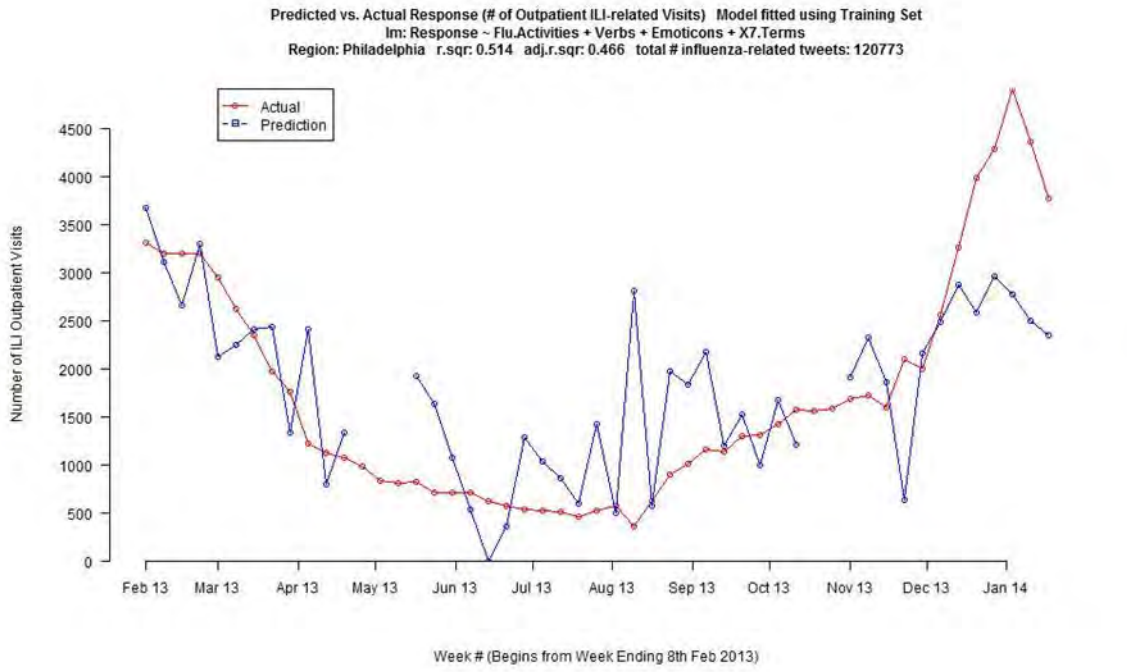

Figure 50. Predicted vs. Actual Number of Outpatient ILI Visits (Philadelphia)

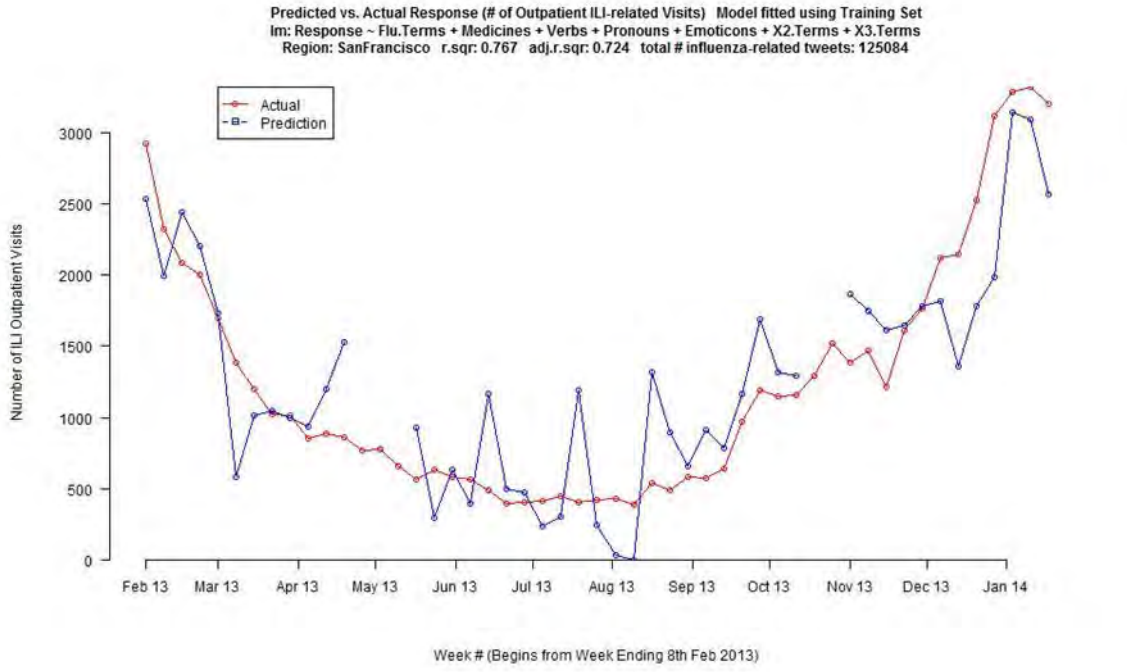

Figure 51. Predicted vs. Actual Number of Outpatient ILI Visits (San Francisco)

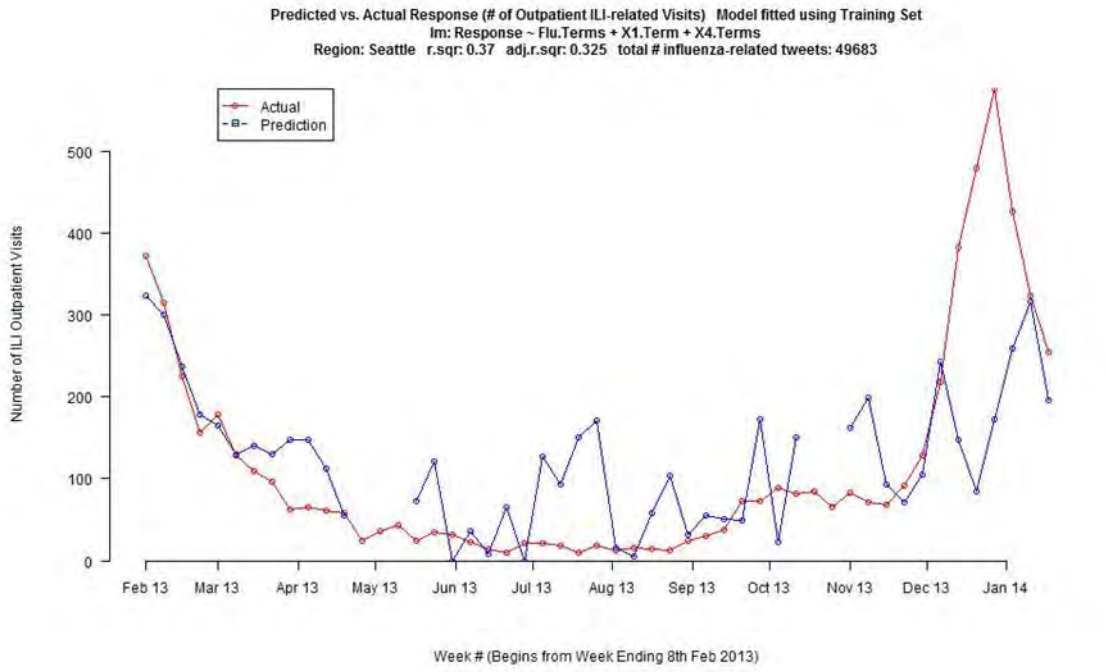

Figure 52. Predicted vs. Actual Number of Outpatient ILI Visits (Seattle)

2. Predicted vs. Actual Number of Collected Respiratory Specimens (Regional)

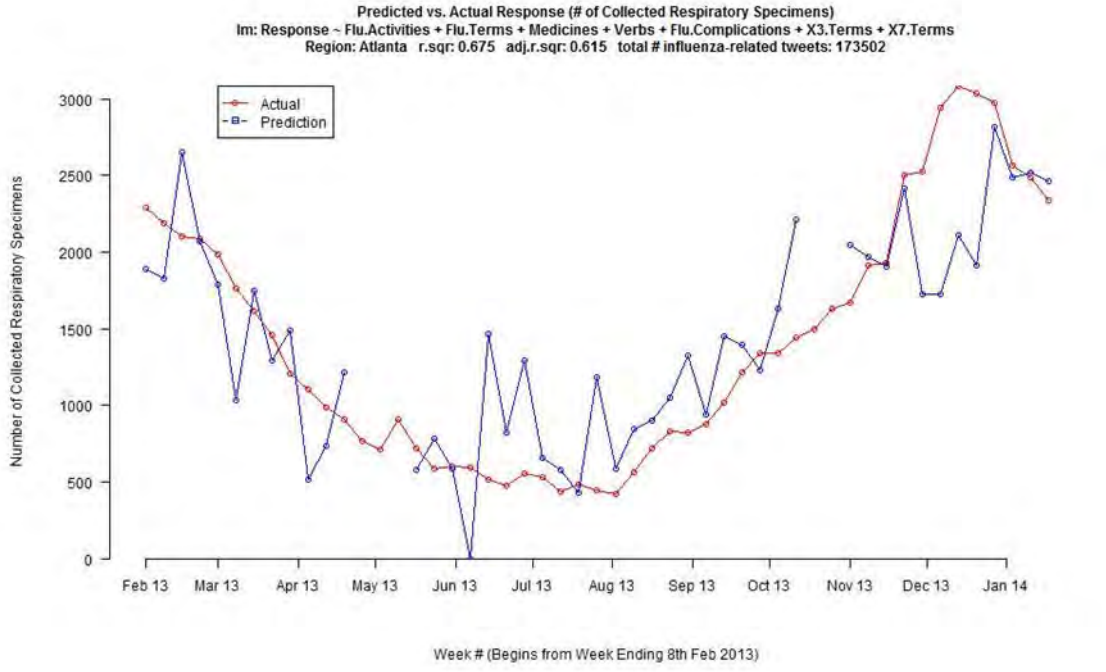

Figure 53. Predicted vs. Actual Number of Collected Respiratory Specimens (Atlanta)

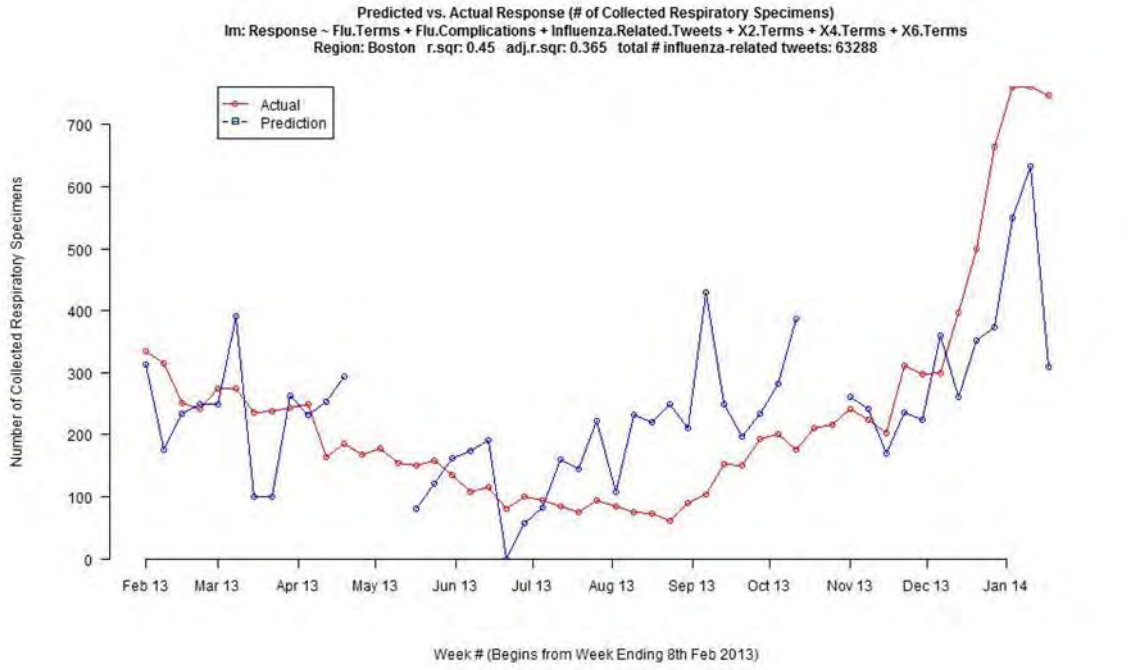

Figure 54. Predicted vs. Actual Number of Collected Respiratory Specimens (Boston)

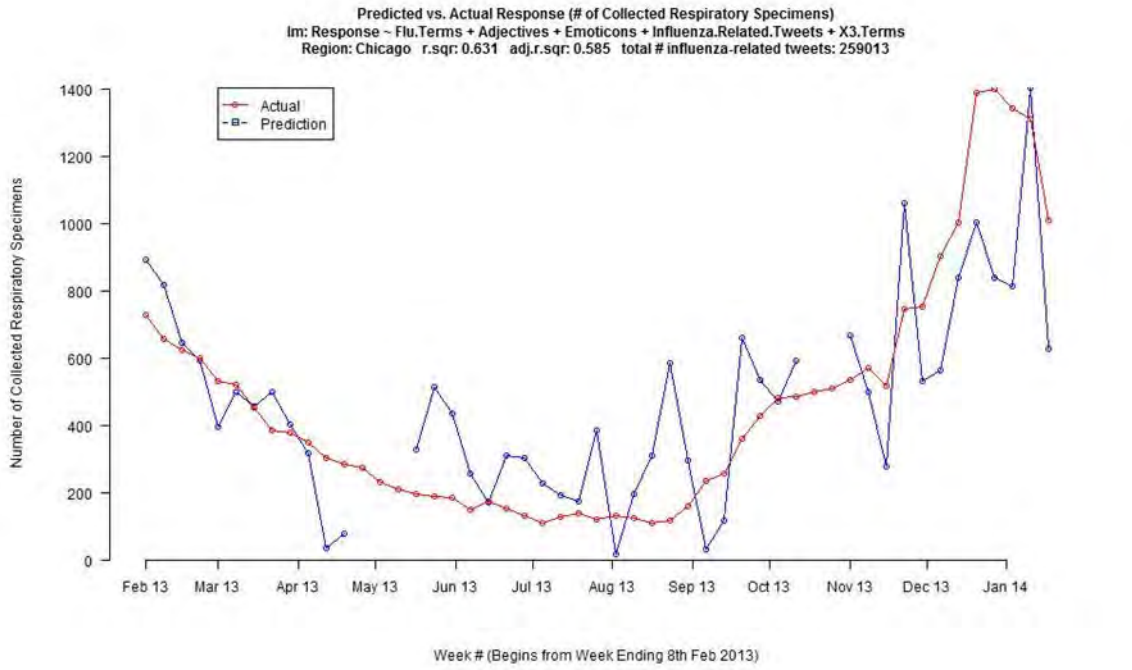

Figure 55. Predicted vs. Actual Number of Collected Respiratory Specimens (Chicago)

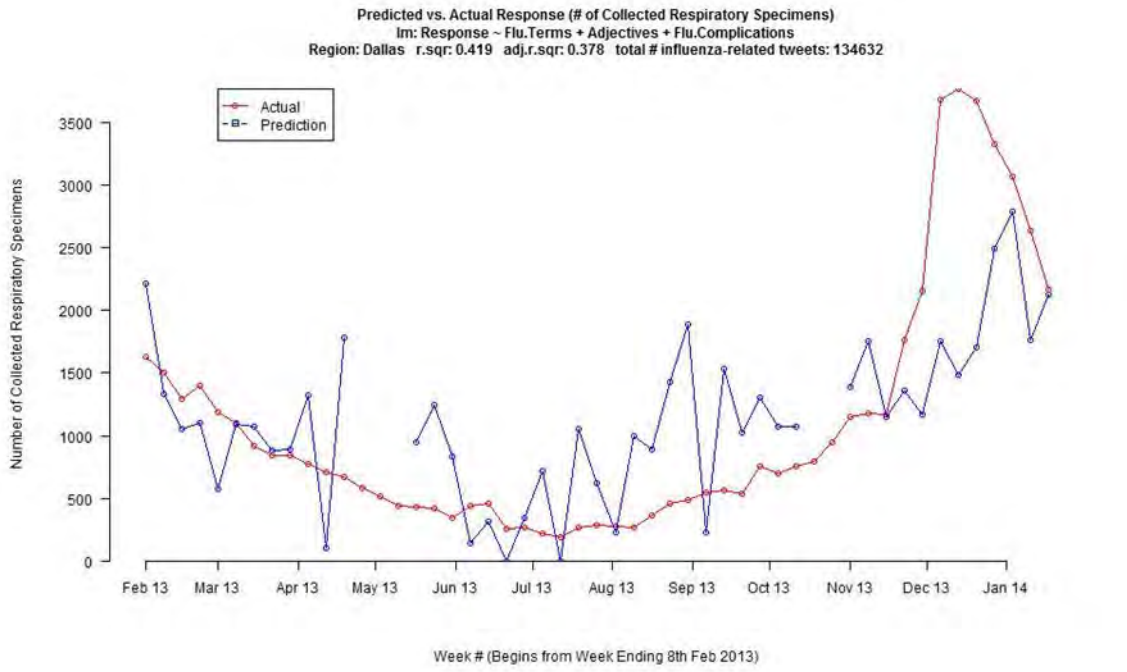

Figure 56. Predicted vs. Actual Number of Collected Respiratory Specimens (Dallas)

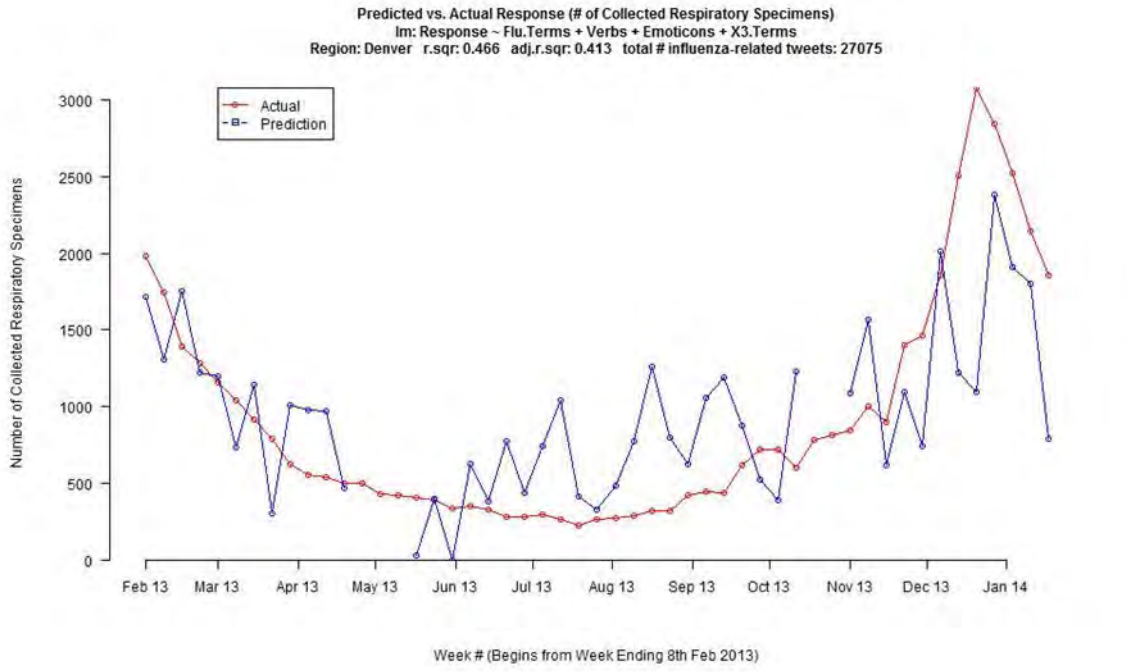

Figure 57. Predicted vs. Actual Number of Collected Respiratory Specimens (Denver)

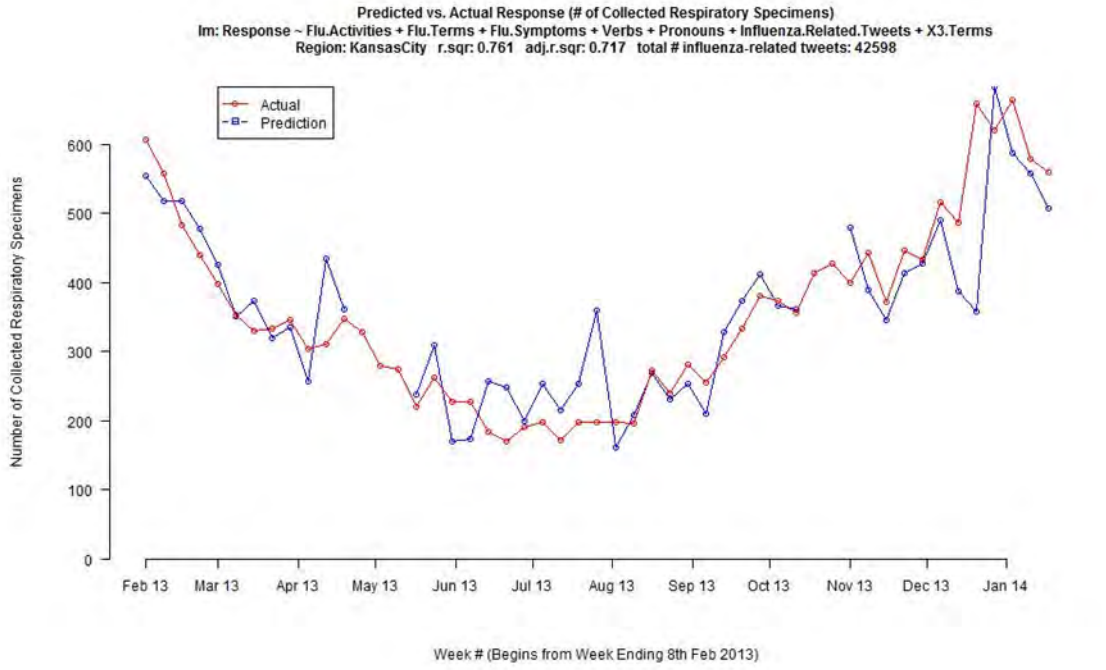

Figure 58. Predicted vs. Actual Number of Collected Respiratory Specimens (Kansas City)

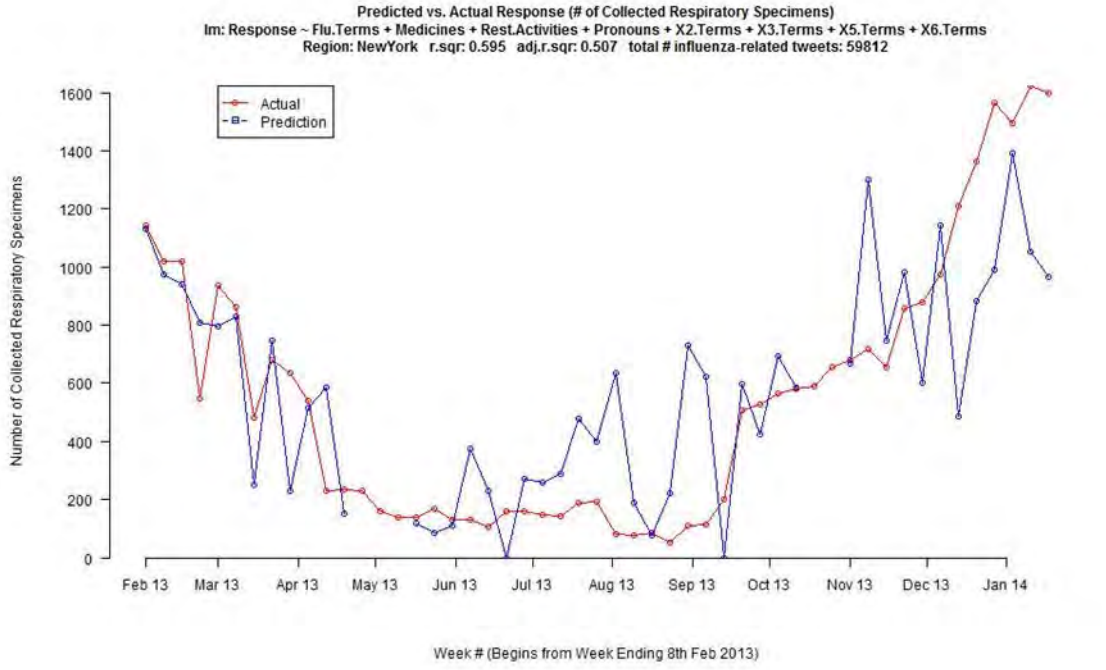

Figure 59. Predicted vs. Actual Number of Collected Respiratory Specimens (New York)

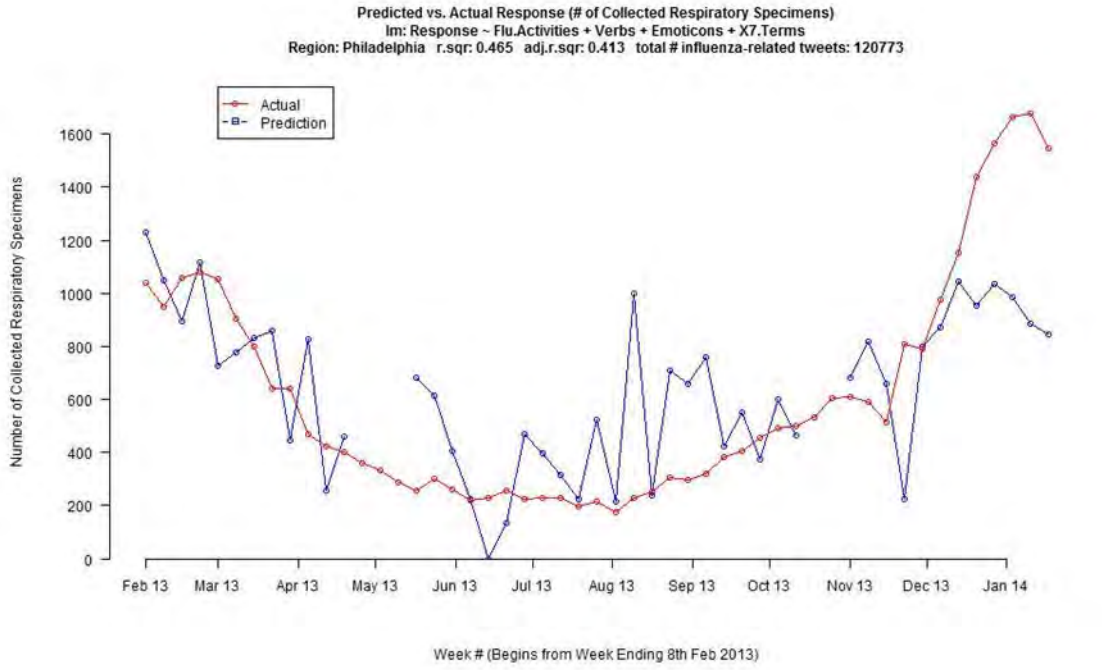

Figure 60. Predicted vs. Actual Number of Collected Respiratory Specimens (Philadelphia)

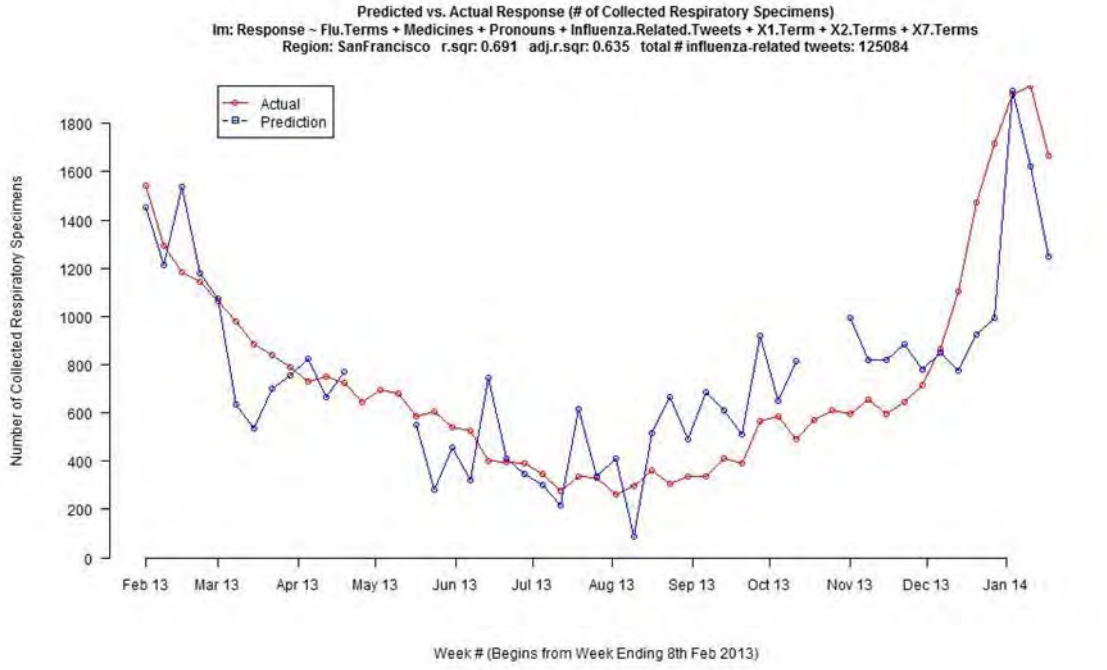

Figure 61. Predicted vs. Actual Number of Collected Respiratory Specimens (San Francisco)

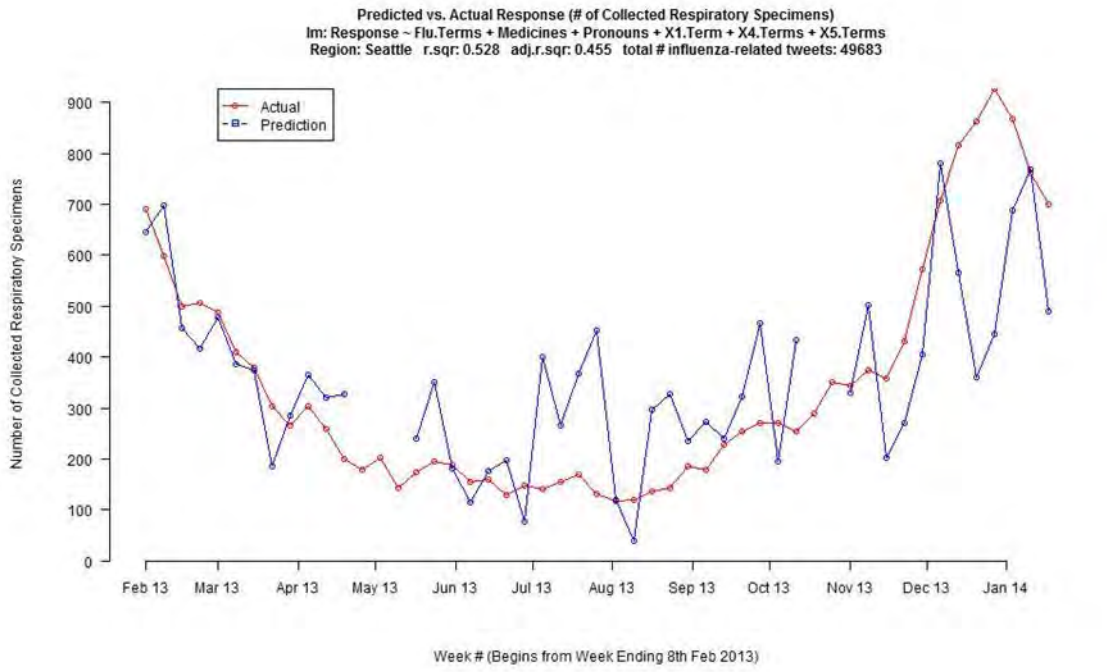

Figure 62. Predicted vs. Actual Number of Collected Respiratory Specimens (Seattle)

3. Predicted vs. Actual Number of Respiratory Specimens Tested Positive for Influenza Type A or B (Regional)

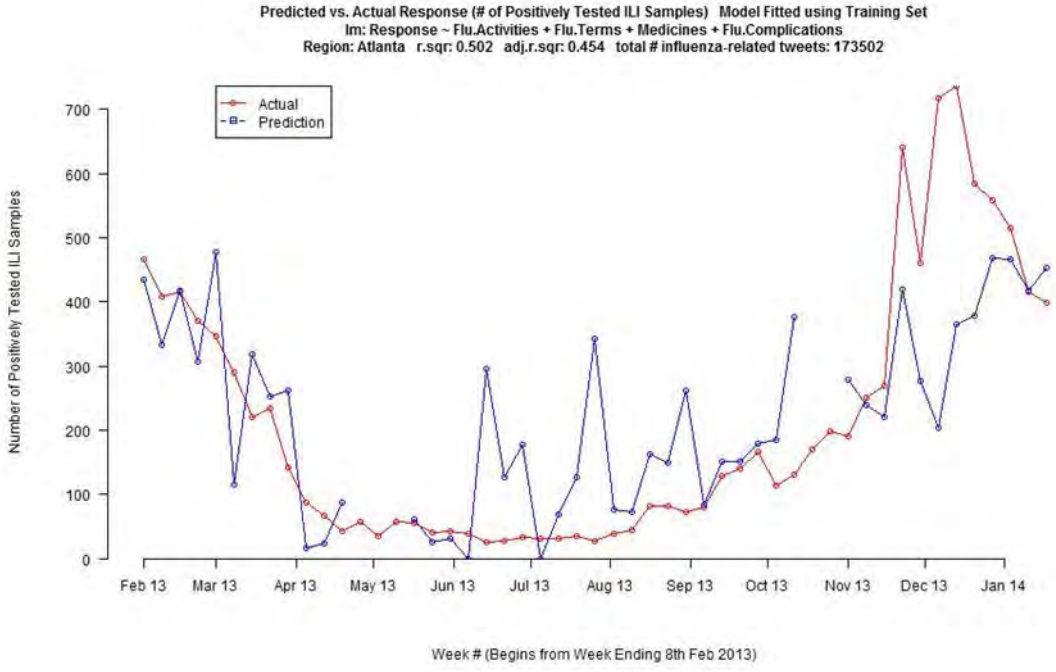

Figure 63. Predicted vs. Actual Number of Respiratory Specimens Tested Positive for Influenza Type A or B (Atlanta)

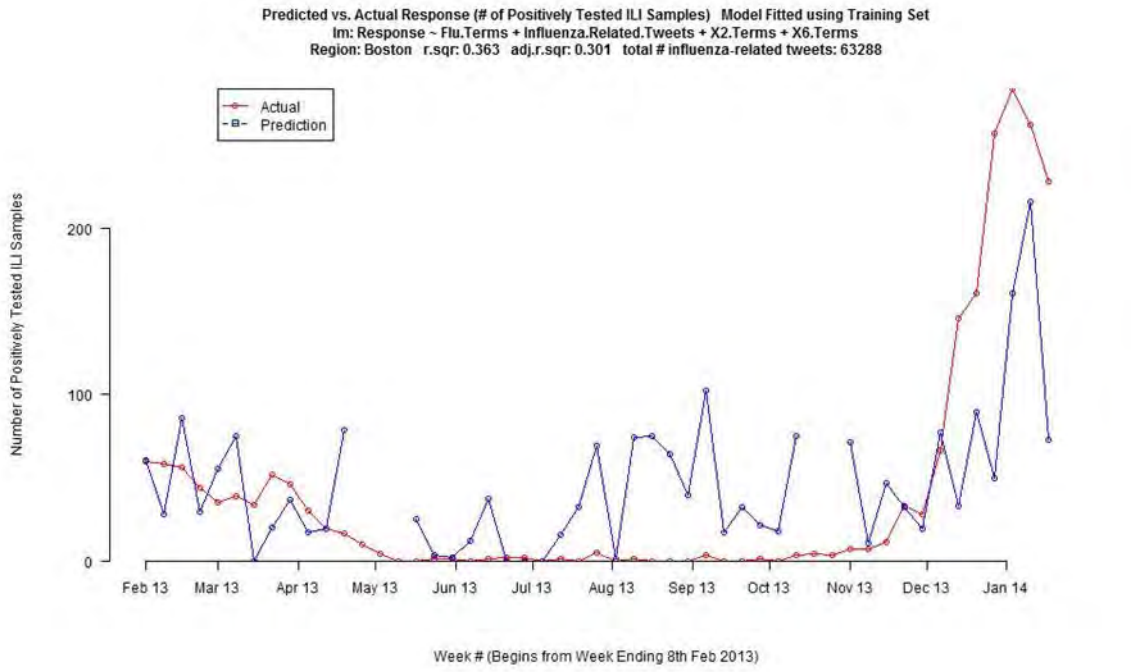

Figure 64. Predicted vs. Actual Number of Respiratory Specimens Tested Positive for Influenza Type A or B (Boston)

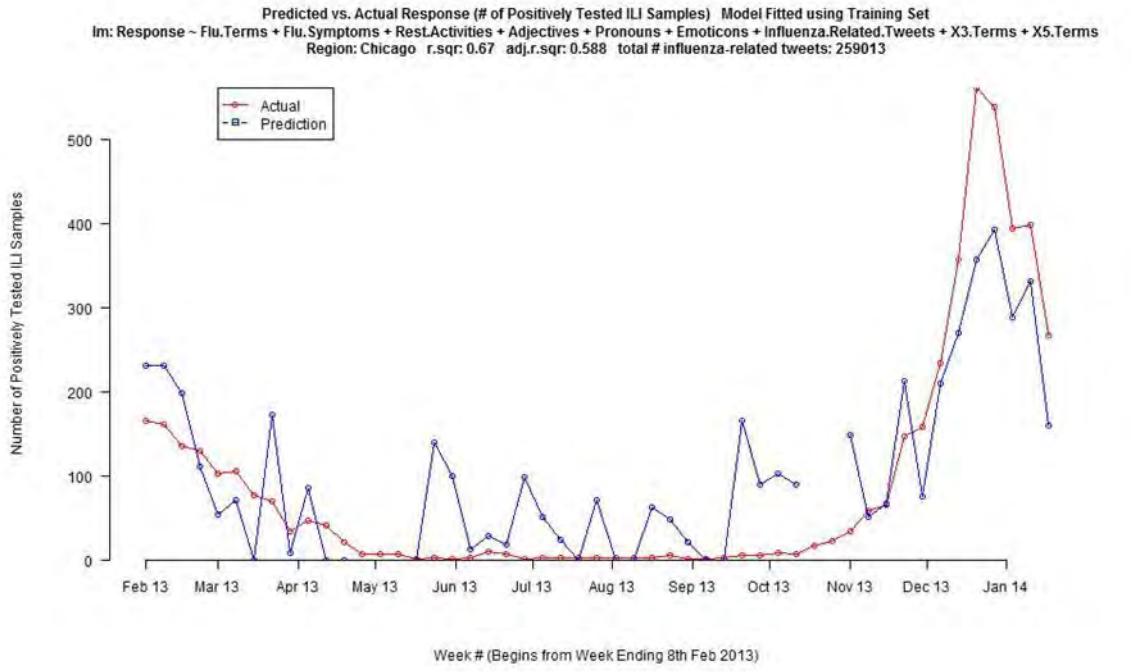

Figure 65. Predicted vs. Actual Number of Respiratory Specimens Tested Positive for Influenza Type A or B (Chicago)

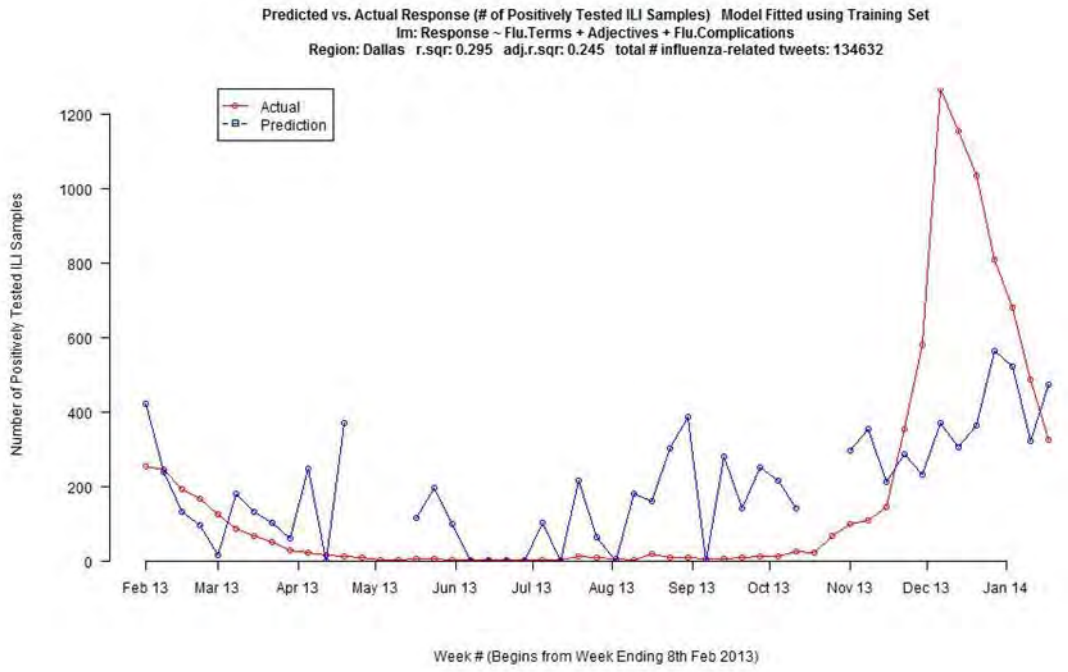

Figure 66. Predicted vs. Actual Number of Respiratory Specimens Tested Positive for Influenza Type A or B (Dallas)

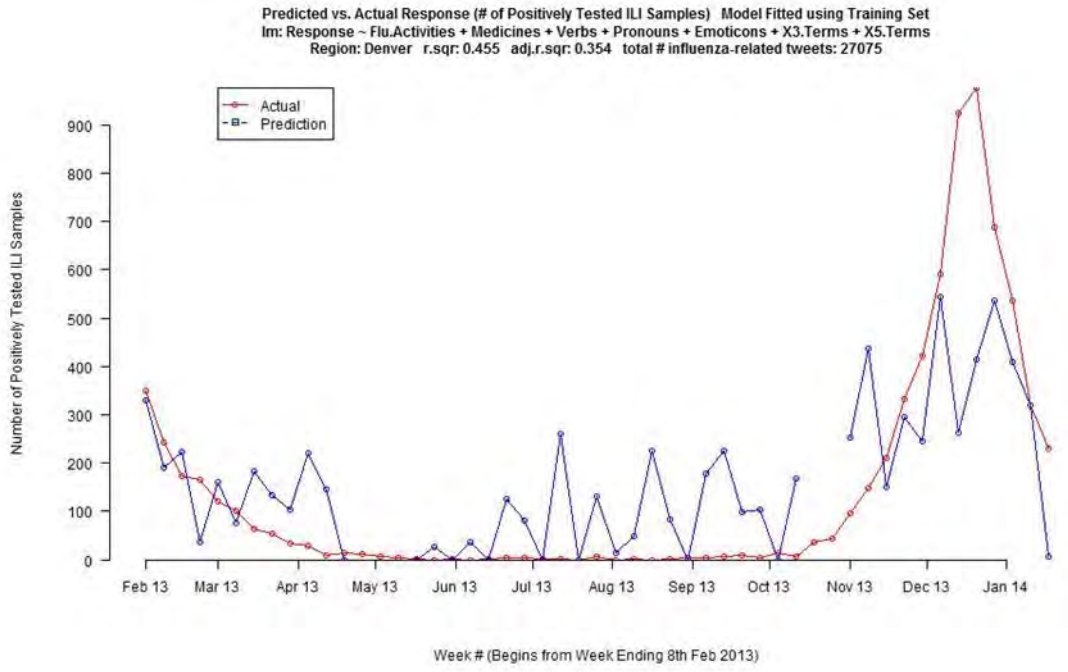

Figure 67. Predicted vs. Actual Number of Respiratory Specimens Tested Positive for Influenza Type A or B (Denver)

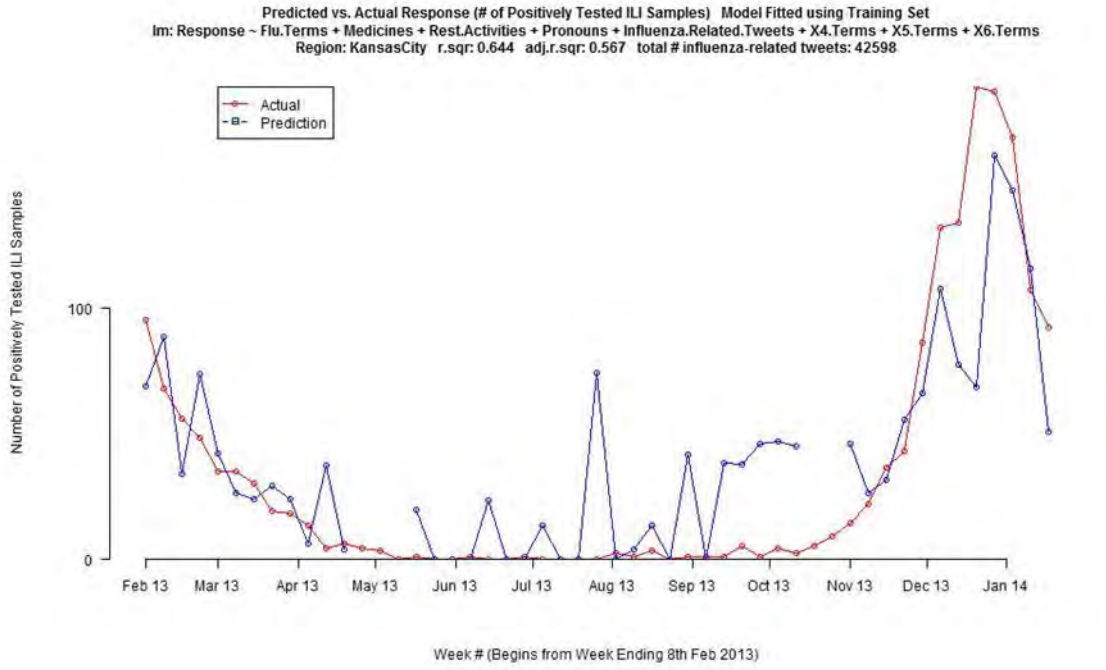

Figure 68. Predicted vs. Actual Number of Respiratory Specimens Tested Positive for Influenza Type A or B (Kansas City)

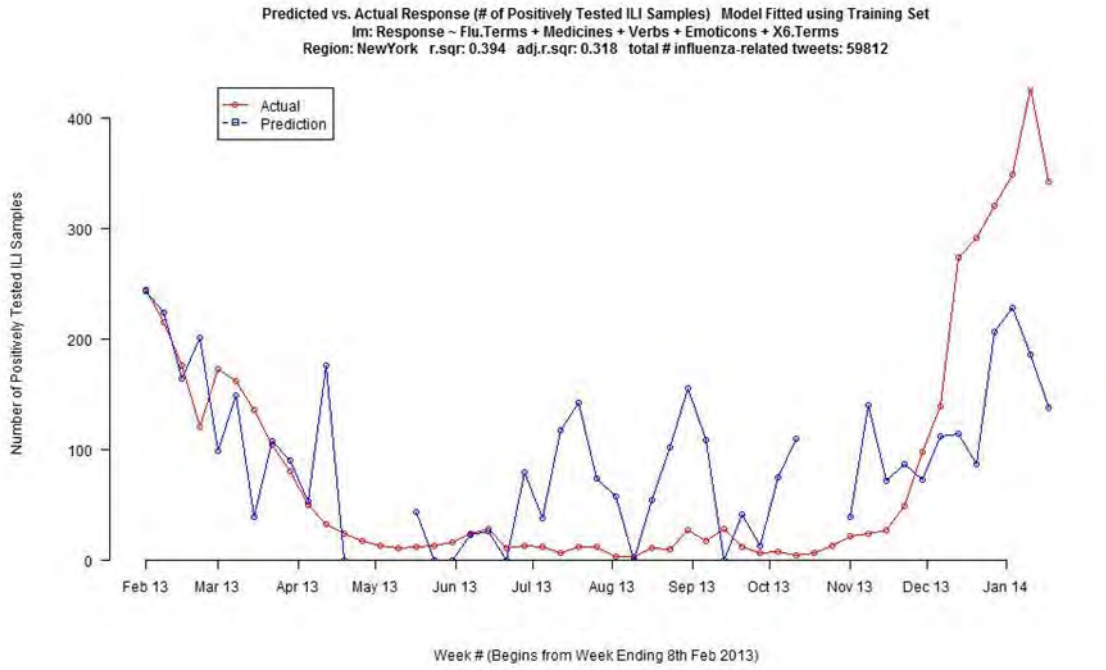

Figure 69. Predicted vs. Actual Number of Respiratory Specimens Tested Positive for Influenza Type A or B (New York)

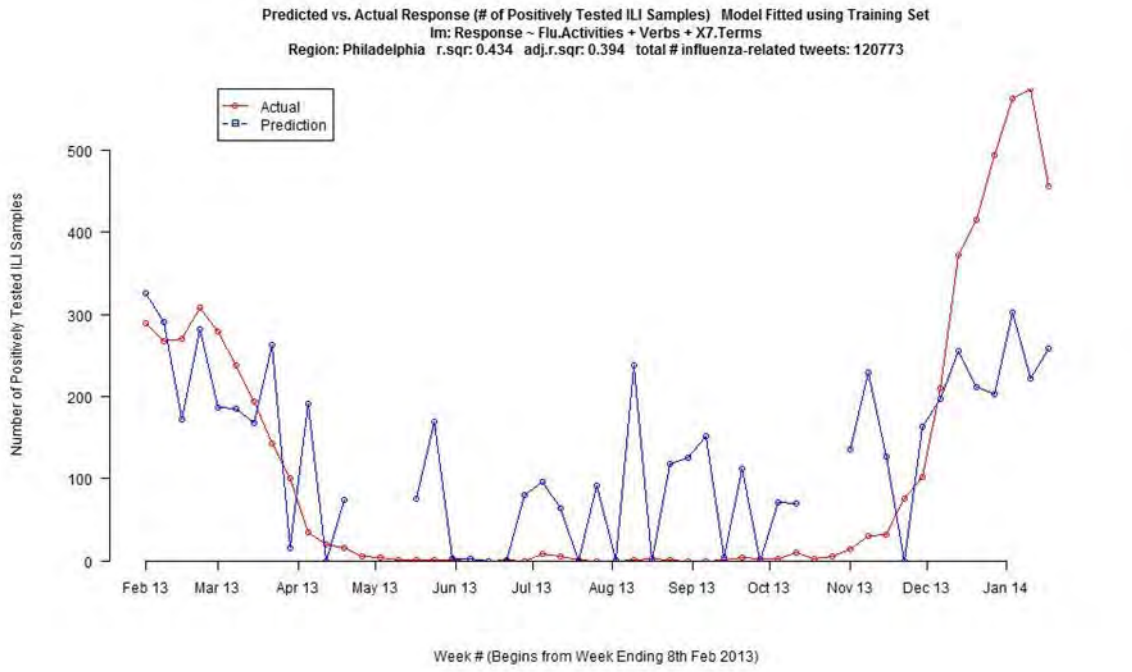

Figure 70. Predicted vs. Actual Number of Respiratory Specimens Tested Positive for Influenza Type A or B (Philadelphia)

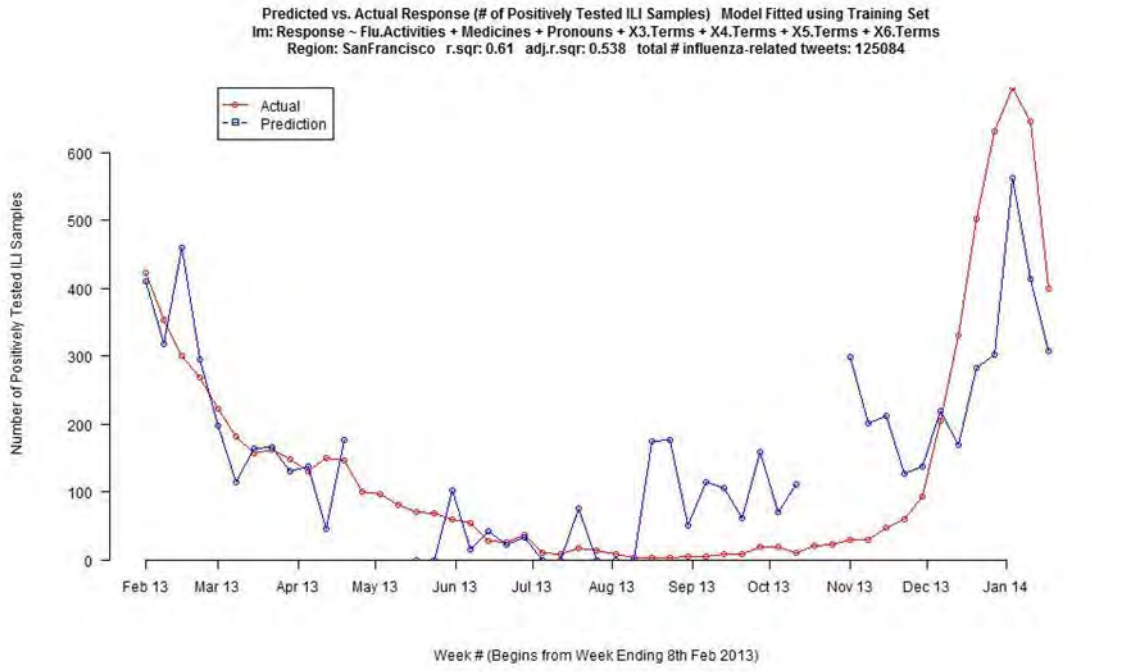

Figure 71. Predicted vs. Actual Number of Respiratory Specimens Tested Positive for Influenza Type A or B (San Francisco)

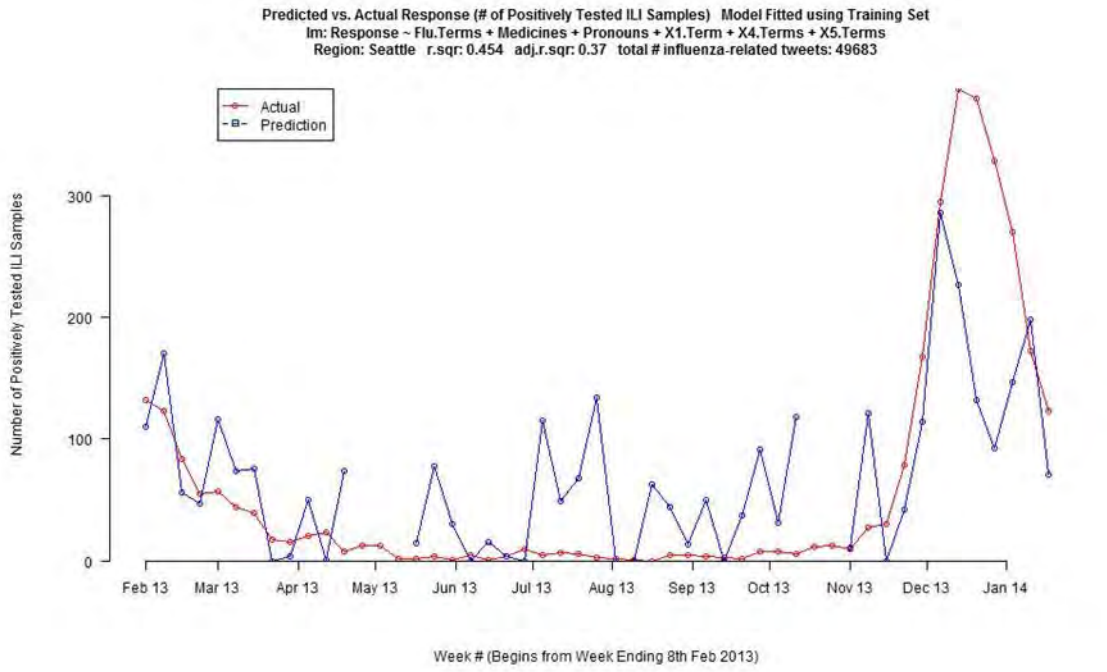

Figure 72. Predicted vs. Actual Number of Respiratory Specimens Tested Positive for Influenza Type A or B (Seattle)

4. Predicted vs. Actual Rate of Influenza-Associated Hospitalizations per 100,000 Population

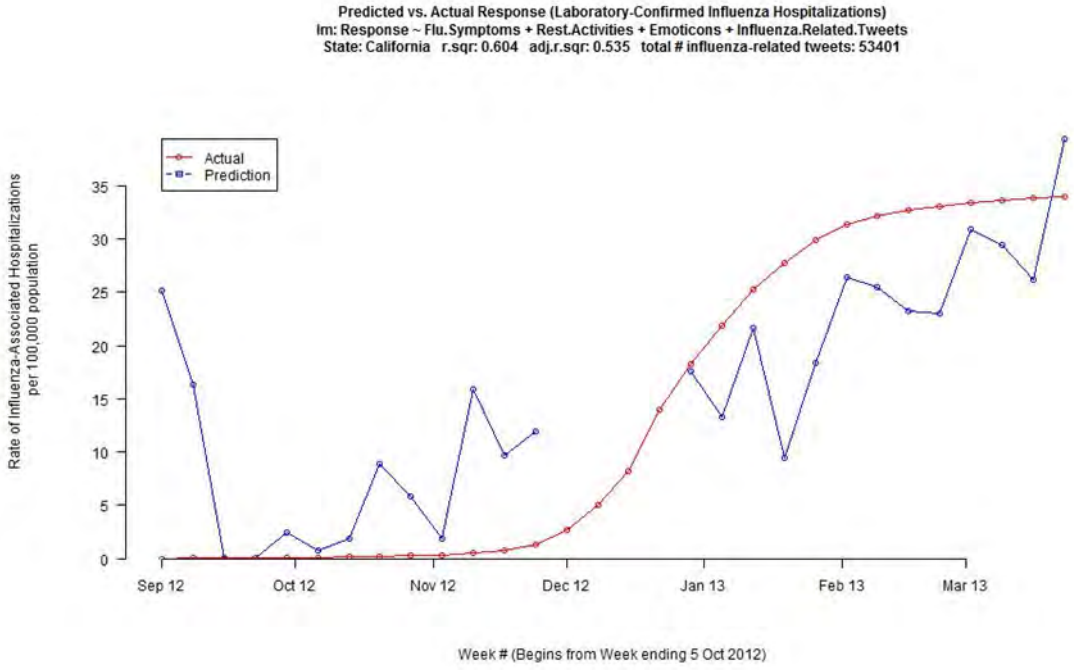

Figure 73. Predicted vs. Actual Rate of Influenza-Associated Hospitalizations per 100,000 Population (California)

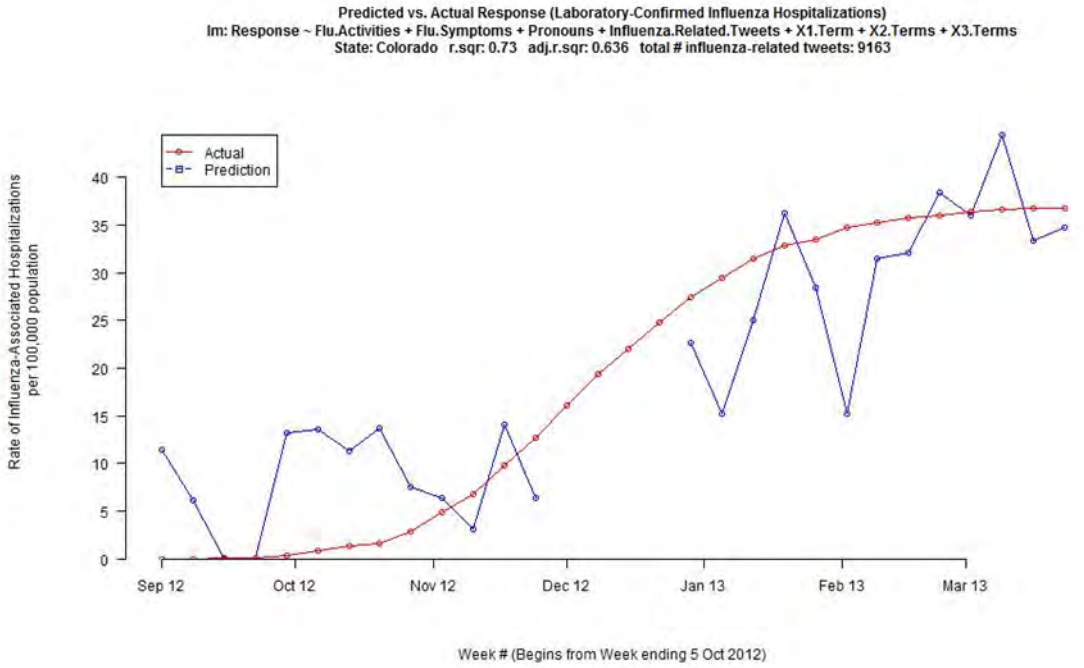

Figure 74. Predicted vs. Actual Rate of Influenza-Associated Hospitalizations per 100,000 Population (Colorado)

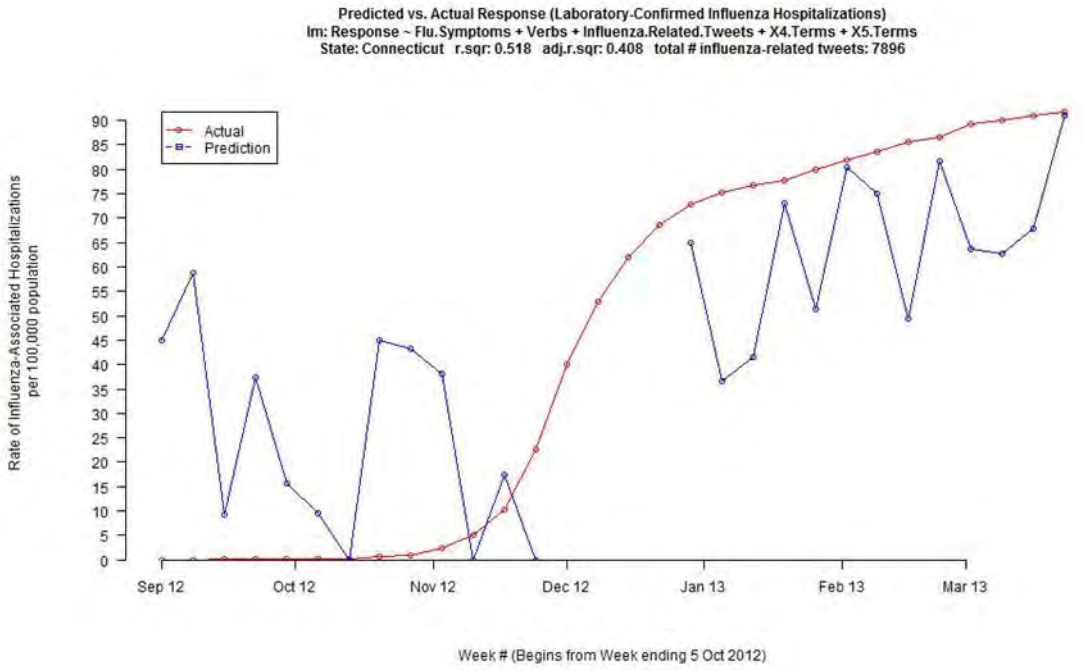

Figure 75. Predicted vs. Actual Rate of Influenza-Associated Hospitalizations per 100,000 Population (Connecticut)

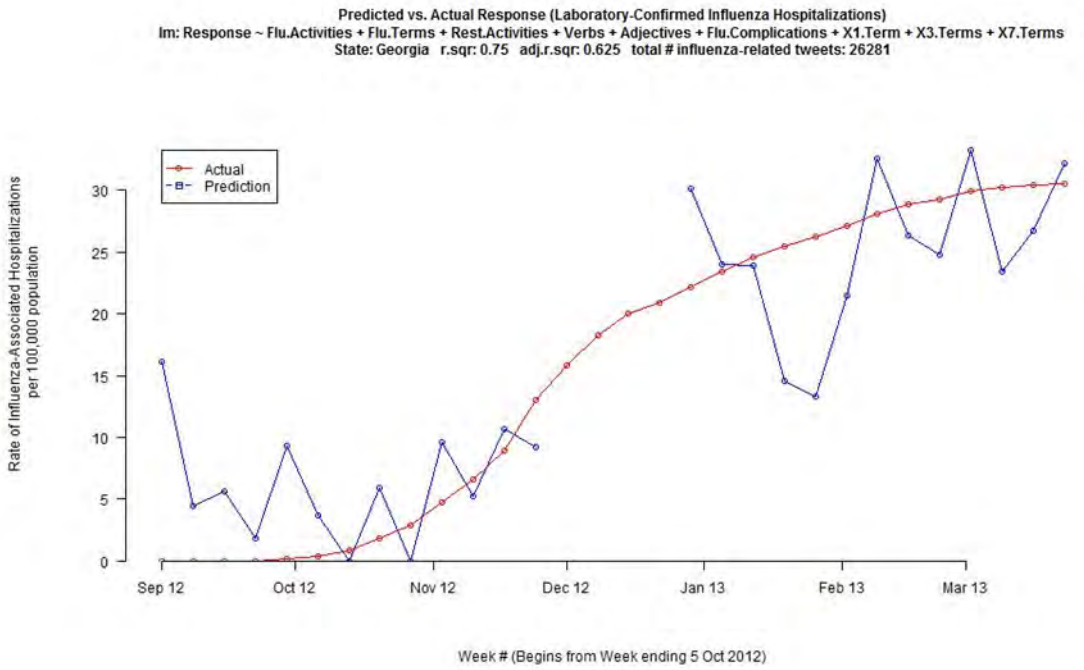

Figure 76. Predicted vs. Actual Rate of Influenza-Associated Hospitalizations per 100,000 Population (Georgia)

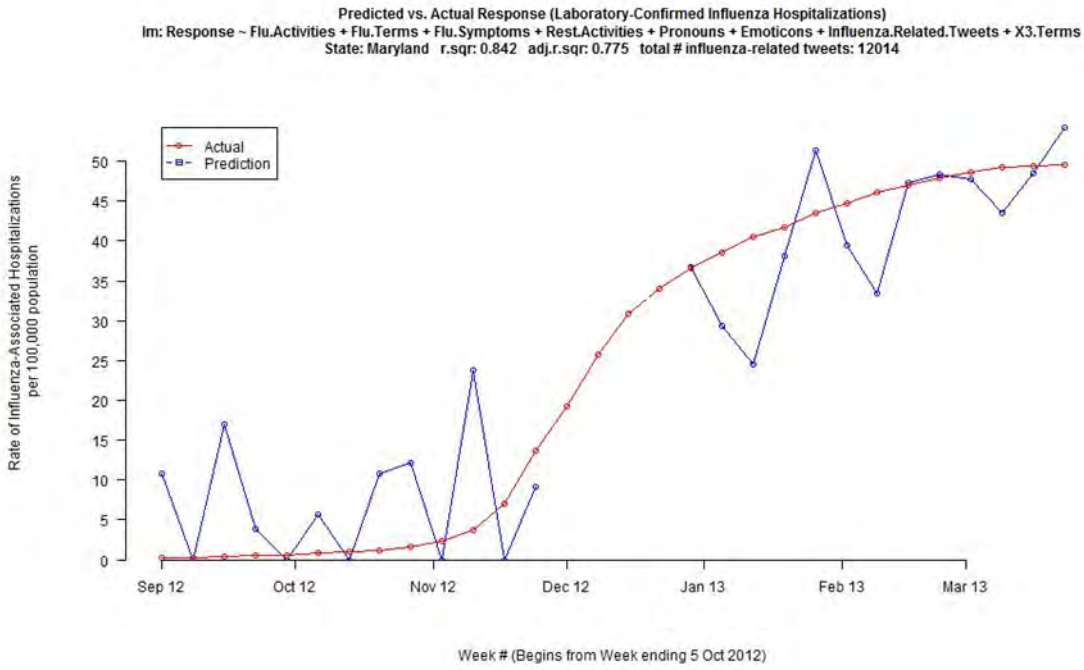

Figure 77. Predicted vs. Actual Rate of Influenza-Associated Hospitalizations per 100,000 Population (Maryland)

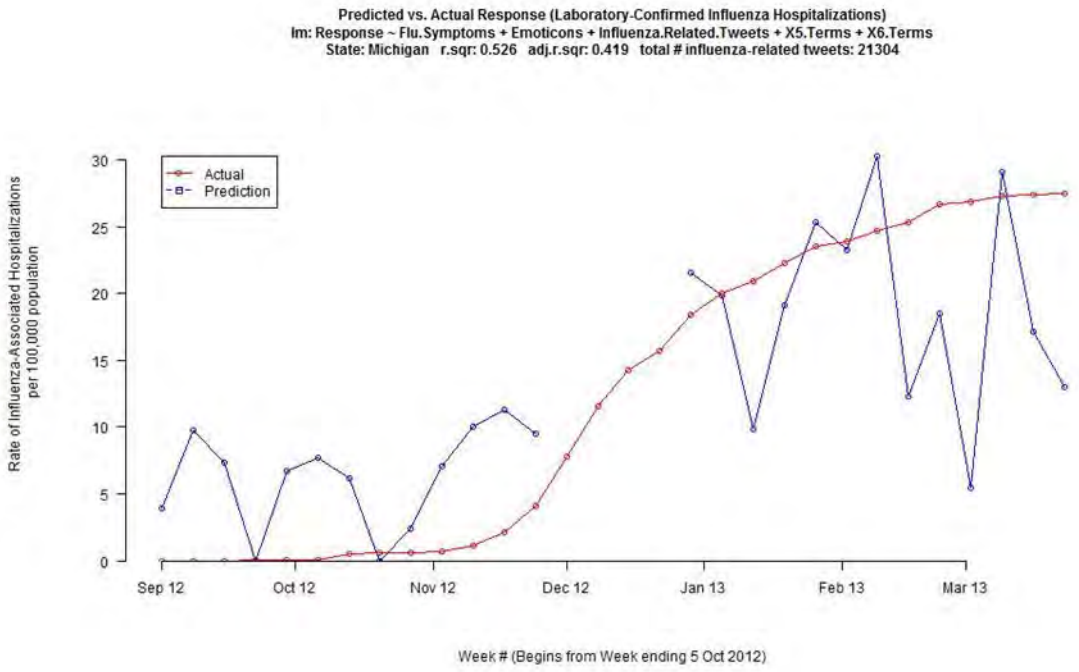

Figure 78. Predicted vs. Actual Rate of Influenza-Associated Hospitalizations per 100,000 Population (Michigan)

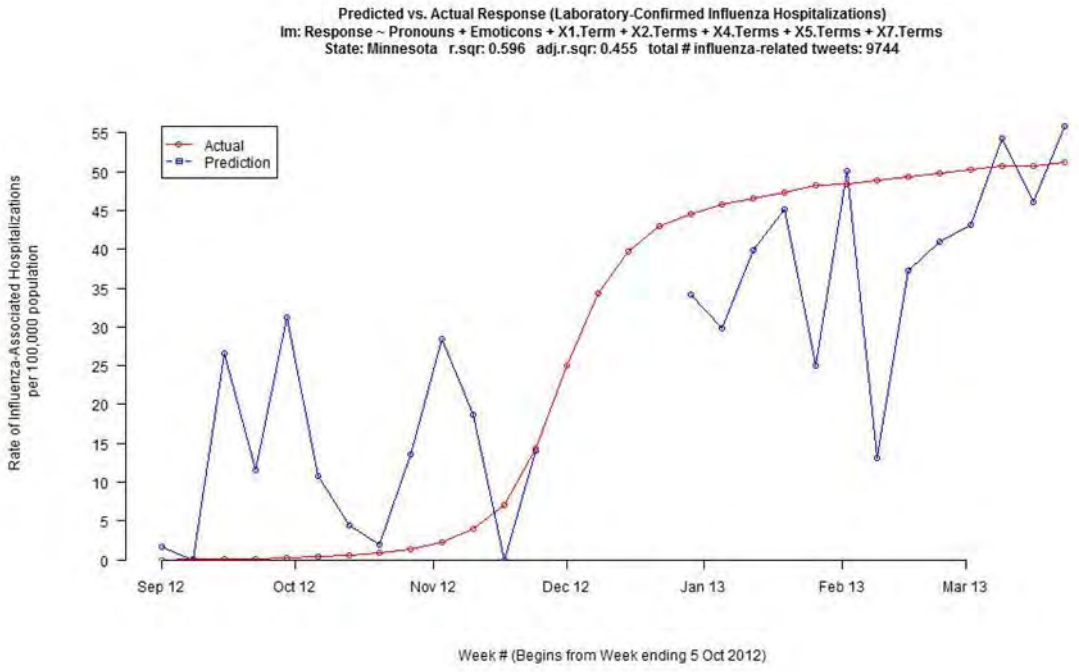

Figure 79. Predicted vs. Actual Rate of Influenza-Associated Hospitalizations per 100,000 Population (Minnesota)

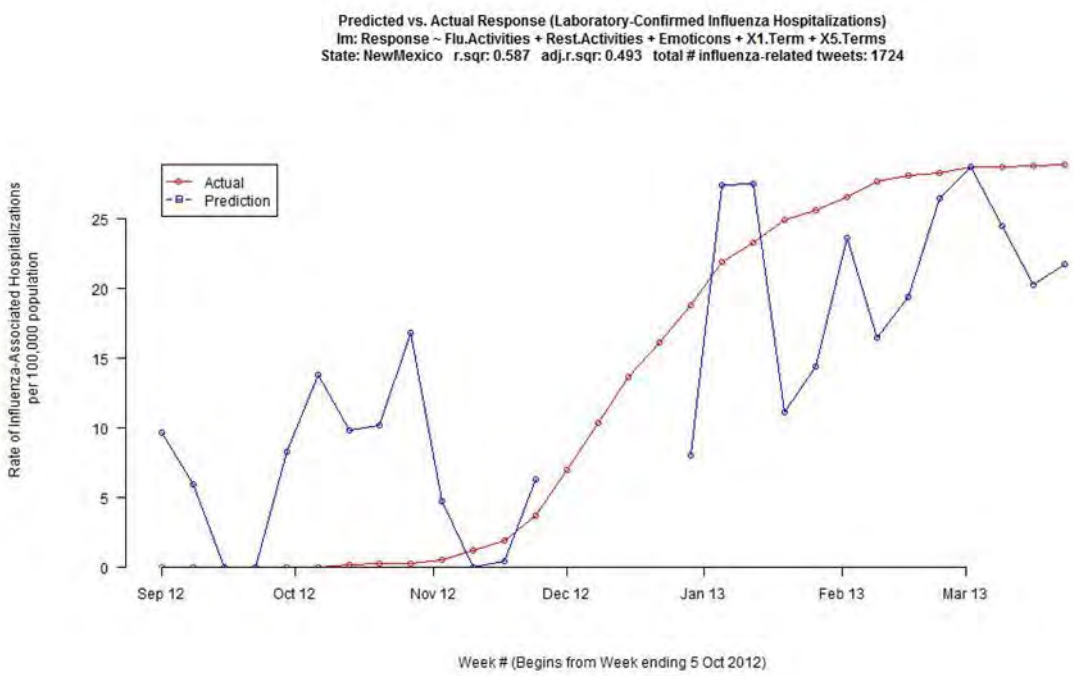

Figure 80. Predicted vs. Actual Rate of Influenza-Associated Hospitalizations per 100,000 Population (New Mexico)

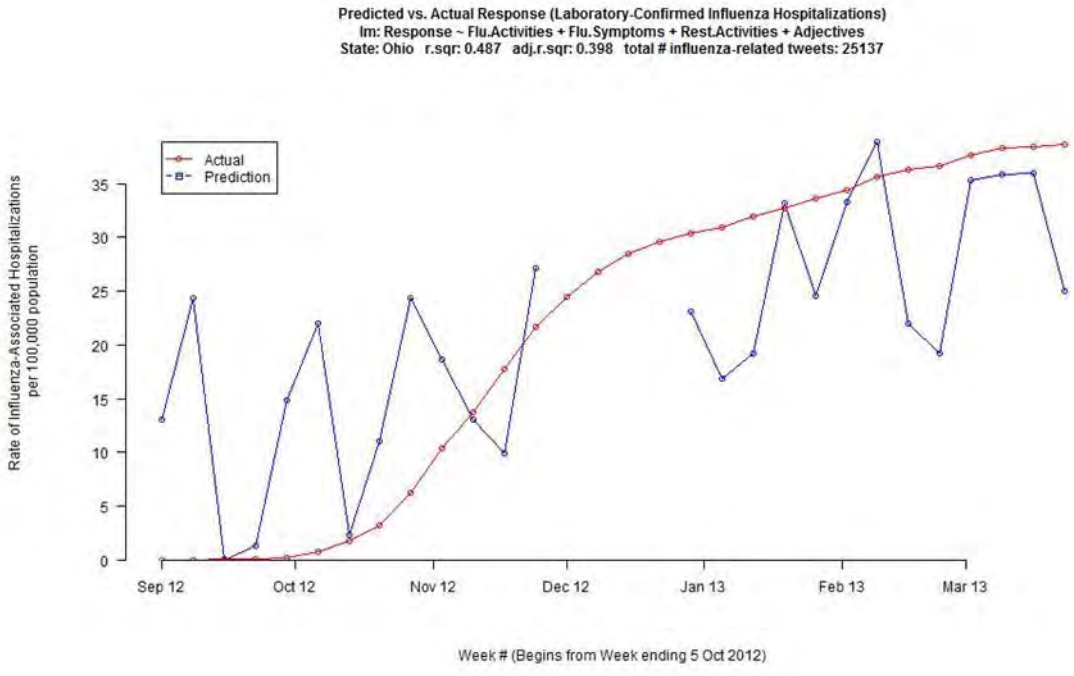

Figure 81. Predicted vs. Actual Rate of Influenza-Associated Hospitalizations per 100,000 Population (Ohio)

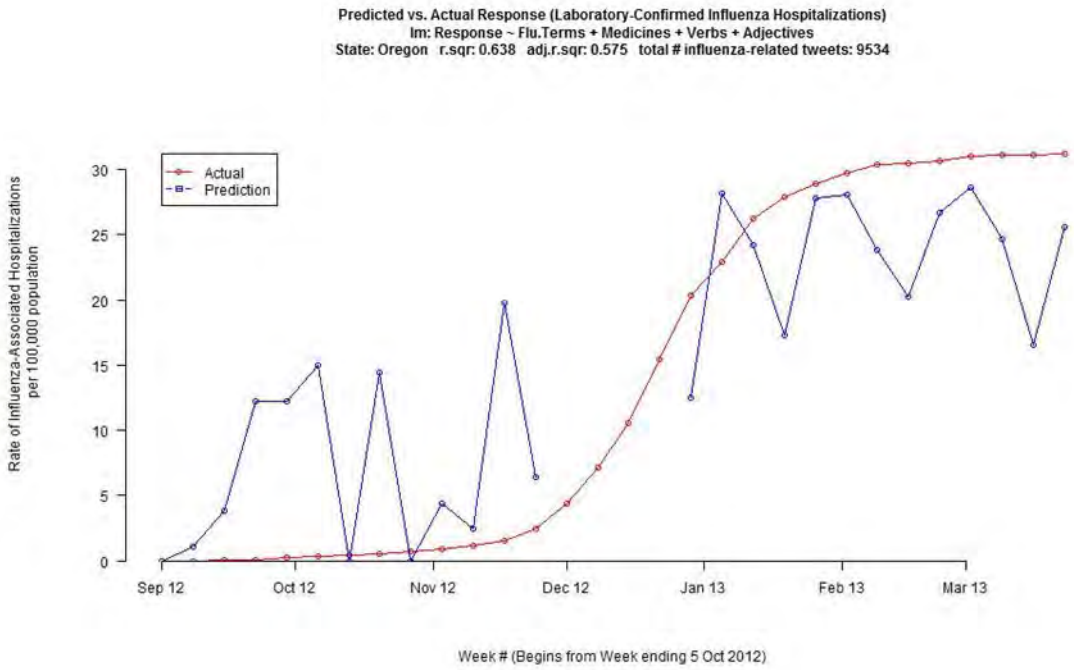

Figure 82. Predicted vs. Actual Rate of Influenza-Associated Hospitalizations per 100,000 Population (Oregon)

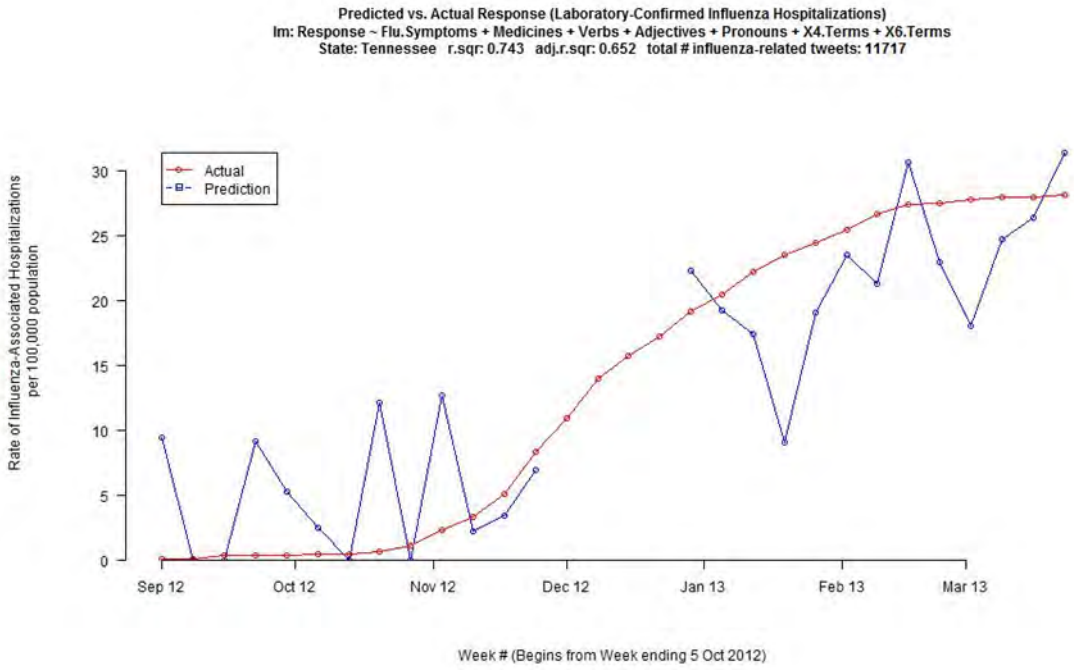

Figure 83. Predicted vs. Actual Rate of Influenza-Associated Hospitalizations per 100,000 Population (Tennessee)

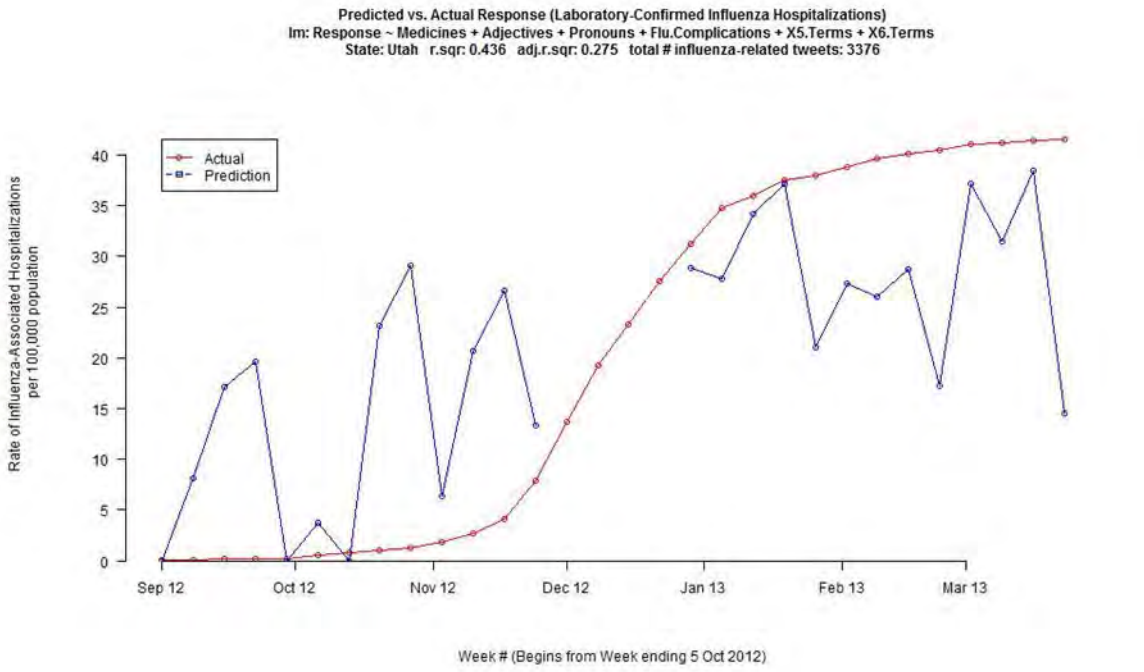

Figure 84. Predicted vs. Actual Rate of Influenza-Associated Hospitalizations per 100,000 Population (Utah)

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX B. DATA

1. List of Terms for Indicative Predictors

Flu.Activities	Flu.Terms	Flu.Symptoms	Medicines	Flu.Complications
Doctor	Flu	Chesty	Medicine	Pneumonia
Clinic	Influenza	Chill	Tylenol	Bronchitis
Hospital	H1N1	Sore Throat	Vicks	Sinus Infection
Doc	Swine	Stuffy Nose	Aspirin	Ear Infection
Hosp	Virus	Fatigue	Medication	
Blood Sample	Viral Infection	Vomiting	Tamiflu	
Blood Test	H3N2	Diarrhoea	Dosage	
Vaccine	Disease	Cough	Dose	
Vaccination		Fever	Treatment	
Pharmacy		Runny Nose	Drugs	
Pharmacies		Aches	Oseltamivir	
Flu Jab		Sick	Prescription	
Flu Shot		Shortness of Breath	Remedy	
		Dizziness	Meds	
		Dizzy	Med	
		Breathless	Dosing	
		Short of Breath		
		Running a Temperature		
		Symptom		
		Body Aches		

Table 14. List of Terms for Each Indicative Predictor Variable

2. List of Terms for Supportive Predictors

Rest.Activities	Verbs	Adjectives	Pronouns	Emoticons
Medical Certificate	Diagnose	Drowsy	I	:’-(
Need Some Rest	Got	Frail	You	:’(
Under the Weather	Down	Bedridden	He	>:[
Off Today	Hit	Unwell	She	:-(
Off Day	Under	Weak	Let	:{
Day Off	Collect	Sickly	It	:[
Taking Off	Recover	Worse	Me	:-[
Time Off	Get	Worst	Him	:<
Need Rest	Contract	Positive	Her	:-<
Need Your Rest	Hospitalize	Antiviral	We	:c
Need A Rest	Admit	Better		:-c
Need To Rest	Get Well	Severe		:(
Take Time Off	Suffer	Persistent		
	Went	Nauseous		
	Bring	Bad		
	Go	Viral		
	Feel	Ill		
	Prescribe			
	Shake Off			
	Ward			
	Vomit			
	Hospitalize			
	Visit			
	Vaccinate			
	Go Away			

Table 15. List of Terms for Each Supportive Predictor Variables

APPENDIX C. SOFTWARE

This section discusses the use of regular expressions (REGEX) for the matching of keywords that exist in the tweets. For this study, matching of keywords is performed to determine the location of the tweet and also to aggregate frequencies of matching influenza-related keywords. Each tweet can contain alphabetical characters, punctuations and frequently used symbols such as the hashtag (#) and emoticons.

REGEX is widely recognizable by various software development environments such as Java, .Net and R. Each REGEX is a sequence of characters that describes a search pattern. Grep is a UNIX command-line function that searches data files that contain a specified REGEX. Similarly, R has a grep function that identifies matching REGEX(s) by searching a given character vector and returns a vector of the indices of the elements of the character vector that yielded a match.

The R function strcount in Figure 85 originates from Madouasse (2012) and is used in this study. The strcount function takes in a pattern and a complete text message, x as arguments and returns the number of pattern occurrences. The function first splits the text message based on the specified split character, into a vector of multiple elements. Next, the grep function is applied to each element in the vector, to determine if any of the elements contains the sequence of characters that matches the pattern.

```
strcount <- function(x, pattern, split){

  unlist(lapply(
    strsplit(x, split),
    function(z) na.omit(length(grep(pattern, z)))
  ))
}
```

Figure 85. R Function: strcount (from Madouasse 2012)

(1) Matching Keywords

The R code snippet in Figure 86 shows an example of determining the number of pattern occurrences for the Flu.Terms keyword category in the message. Flu.Terms

keywords are basically terms that are directly related to flu such as flu, influenza and H1N1.

As shown in this example, the REGEX is always expressed as a string. In each REGEX, vertical bars can be used to separate alternative keywords. Hence, by using the specified REGEX in this example, a match exists if any of the alternative keywords are found in the message.

Prior to the strcount function call, punctuations are first removed from the message. This is to ensure that no punctuations are attached (prefixed or suffixed) to the individual words in the message vector after the strsplit function is called. As the grep function only matches the exact patterns defined in the REGEX, the existence of an attached punctuation to a match keyword will result in a no-match.

```
message.remove.punc <- gsub('[[:punct:]]','',message)
message.vector <- strsplit(message.remove.punc, " ")
count_FT <- strcount(
  tolower(message.vector), "\\b(flu)\\b|influenza|h1n1|swine|virus|h3h2|disease", " ")
```

Figure 86. R Code Snippet: Matching Keywords

(2) Matching Key Phrases

For keywords that exist in the form of a phrase (e.g., viral infection), the input message and pattern for the strcount function need to be changed. The two changes are for: (1) message argument: Pass the message with no empty spaces used between the words in the message (2) pattern argument: define the REGEX by combining the words in the phrase into a single term (e.g., viralinfection). Figure 87 shows the R code snippet for matching key phrases.

```
message.remove.space <- gsub('[[:space:]]','',message)
count_FT <- count_FT +
  strcount(tolower(message.remove.space), "viralinfection", " ")
```

Figure 87. R Code Snippet: Matching Key Phrases

(3) Matching Pronouns

The R grep function basically returns the count of matching patterns defined in the REGEX. For a REGEX that is specified as “I,” grep will count the number of occurrences of the alphabet “i” in the sentence. Hence, we will need to use the metacharacter “\\b” to ensure that the boundary (first and last character) of a matching word is not a word character [A-Z].

A couple of initial steps are required: (1) add empty spaces to the beginning and end of the message, (2) replace the empty spaces in the message with “-|-,” and (3) strsplit the message using the separator “|.” By performing the three steps, we obtained a vector of elements, where each element is a word in the message that is now prefixed and suffixed by a “-.” Next, for the REGEX declaration, prefix and suffix each pronoun with the metacharacter “\\b” to qualify matching words that begin and end with a non-word character. Figure 88 shows the R code snippet for matching pronouns.

```
message <- paste("", message, "", sep = " ")
message.re <- gsub('[[:space:]]','-|-',message)
message.vector.pronoun <- strsplit(message.re, "|")
count_PRONOUN <- strcount(
  tolower(message.vector), "\\b(i)\\b|\\b(im)\\b|\\b(you)\\b|\\b(we)\\b", " ")
```

Figure 88. R Code Snippet: Matching Pronouns

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF REFERENCES

Aramaki, Eiji, Sachiko Maskawa, and Mizuki Morita. n.d. "Twitter Catches The Flu: Detecting Influenza Epidemics Using Twitter." *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics) 1568–1576.

Babwin, Don. 2009. "Hundreds of Schools Close from H1N1–Los Angeles Times." *The Associated Press*, October 28. Accessed July 10, 2014, http://www.latimes.com/health/sns-health-swine-flu-schools-story.html.

Beevolve. 2012. An Exhaustive Study of Twitter Users Across the World. October 10. Accessed July 27, 2014, http://www.beevolve.com/twitter-statistics/#a3.

Bo, Han, Paul Cook, and Timothy Baldwin. n.d. "A Stacking-based Approach to Twitter User Geolocation Prediction." *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (Association for Computational Linguistics) 7–12.

Butler, Declan. 2013. "When Google Got Flu Wrong." *Nature* 494, no. 7436: 155–156. doi:10.1038/494155a.

Buttrey, Samuel E. 2012. xval: R Function that Does Cross Validation for a Set of Predictor Variables. *Unpublished.* Naval Postgraduate School. Monterey.

Cain, Daniel T. 2013. "Twituational awareness: gaining situational awareness via crowdsourced #disaster." Master's thesis, Naval Postgraduate School, 2013.

CDC. 2014. "A Weekly Influenza Surveillance Report Prepared by the Influenza Division." CDC–Seasonal Influenza (Flu)–Weekly U.S. Map: Influenza Summary Update. July 11. Accessed July 13, 2014, http://www.cdc.gov/flu/weekly/usmap.htm.

———.2014. "Influenza National and Regional Level Graphs and Data." *FLUVIEW: National and Regional Level Outpatient Illnesses and Viral Surveillance.* Accessed July 18, 2014, http://gis.cdc.gov/grasp/fluview/fluportaldashboard.html.

———.2014. "CDC Novel H1N1 Flu|CDC Estimates of 2009 H1N1 Influenza Cases, Hospitalizations and Deaths in the United States." June 24. Accessed July 15, 2014, http://www.cdc.gov/h1n1flu/estimates_2009_h1n1.htm.

———.2014. "Vaccine Effectiveness–How Well Does the Flu Vaccine Work?" May 1. Accessed July 12, 2014, http://www.cdc.gov/flu/about/qa/vaccineeffect.htm.

———.2010. “CDC Novel H1N1 Flu | The 2009 H1N1 Pandemic: Summary Highlights, April 2009–April 2010.” June 16. Accessed August 10, 2014, http://www.cdc.gov/h1n1flu/cdcresponse.htm.

Couture-Beil, Alex. 2014. “rjson: JSON for R, R package version 0.2.14.” Accessed September 16, 2014, http://CRAN.R-project.org/package=rjson.

Culotta, Aron. 2010. “Towards Detecting Influenza Epidemics by Analyzing Twitter Messages.” *Proceedings of the First Workshop on Social Media Analytics* (Association for Computing Machinery) 115–122.

Elson, Sara B, Douglas Yeung, Parisa Roshan, S R Bohandy, and Alireza Nader. 2012. *Using Social Media to Gauge Iranian Public Opinion and Mood After the 2009 Election*. Santa Monica: RAND Corporation.

Evans, Mark. 2010. *Exploring the Use of Twitter around the World: Sysomos Blog.* January 14. Accessed August 11, 2014, http://blog.sysomos.com/2010/01/14/exploring-the-use-of-twitter-around-the-world/.

Ginsberg, Jeremy, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, and Larry Brilliant. 2009. “Detecting Influenza Epidemics Using Search Engine Query Data.” *Nature*. doi:10.1038/nature07634.

Google. 2014. “Data Source: Google Flu Trends.” Accessed August 11, 2014, http://www.google.org/flutrends.

———.n.d. “Google Flu Trends | How.” Accessed August 1, 2014, http://www.google.org/flutrends/about/how.html.

Kim, Eui-Ki, Jong Hyeon Seok, Jang Seok Oh, Hyong Woo Lee, and Kyung Hyun Kim. 2013. “Use of Hangeul Twitter to Track and Predict Human Influenza Infection.” *PLoS ONE* 8, no. 7: e69305. doi:10.1371/journal.pone.0069305.

Koyak, Robert A. 2013. Bestsubxval: R Function that Works in Conjunction with Regsubsets Function and xval Function to Return a Best Subset of Predictors. *Unpublished.* Naval Postgraduate School. Monterey.

Madouasse, Aurélien. 2014. “R code: How to Count the Number of Occurrences of a Substring within a String.” May 24. Accessed June 11, 2014, https://aurelienmadouasse.wordpress.com/2012/05/24/r-code-how-the-to-count-the-number-of-occurrences-of-a-substring-within-a-string/.

R Core Team. 2013. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Project for Statistical Computing.

Stone, Biz. 2009. “Location, Location, Location | Twitter Blogs.” August 20. Accessed September 16, 2014, https://blog.twitter.com/2009/location-location-location.

Lumley, Thomas using Fortran code by Miller, Alan. 2009. “Leaps: Regression Subset Selection. R Package Version 2.9.” Accessed September 16, 2014, http://CRAN.R-project.org/package=leaps.

Twitter. 2014. “About Twitter, Inc.” Accessed August 11, 2014. https://about.twitter.com/company.

U.S. Department of Health & Human Services. 2006. “HHS Region Map.” *HHS.gov*. June 19. Accessed June 15, 2014, http://www.hhs.gov/about/regionmap.html.

WHO. 2014. “WHO | FluNet.” Accessed August 11, 2014, http://www.who.int/influenza/gisrs_laboratory/flunet/en/.

Wikipedia Contributors. 2014. “List of Emoticons.” June 21. Accessed July 27, 2014, http://en.wikipedia.org/w/index.php?title=List_of_emoticons&oldid=613828247.

Yuan, Qingyu, Elaine O. Nsoesie, Benfu Lv, Geng Peng, and Rumi Chunara. 2013. “Monitoring Influenza Epidemics in China with Search Query from Baidu.” *PLoS ONE* 8, no. 5: e64323. doi:10.1371/journal.pone.0064323.

THIS PAGE INTENTIONALLY LEFT BLANK

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California